Latent class trees

van den Bergh, Mattis

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
van den Bergh, M. (2018). *Latent class trees*. Proefschriftenmaken.nl.

Latent Class Trees

Mattis van den Bergh
Tilburg University

Printing was financially supported by Tilburg University.

ISBN:	978-94-6295-790-9
Printed by:	ProefschriftMaken | | www.proefschriftmaken.nl
Cover design:	Casper Harinck

Latent Class Trees

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan Tilburg University,
op gezag van de rector magnificus, prof. dr. E.H.L. Aarts,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie in
de aula van de Universiteit

op vrijdag 5 januari 2018 om 14.00 uur

door

Mattis van den Bergh

geboren op 28 mei 1988
te Amsterdam

Promotor:	prof. dr. J.K. Vermunt
Copromotor:	dr. V.D. Schmittmann
Overige leden van de Promotiecommissie:	prof. dr. C.V. Dolan
	prof. dr. J.P.T.M. Gelissen
	dr. S.K. Johnson
	dr. E.M.L. Dusseldorp
	dr. D.W. van der Palm

In loving memory of Janneke de Kort

- Ik mis je -

Contents

Chapter 1

Introduction

1.1 The Latent Class model

The last decades, Latent Class (LC) analysis has become a popular tool among social and behavioral scientists, as often substantive concepts of the research at hand cannot be measured directly. LC analysis allows to detect unobserved homogeneous subgroups in multivariate categorical data that can be interpreted substantively. For example, using a LC analysis Savage et al. (2013) identified different social classes using a set of questions on social, cultural, and economic topics, Jansen and van der Maas (1997) found different developmental stages in children based on rule assessment of the balance scale task (Inhelder & Piaget, 1958; Siegler, 1976), and Mulder, Vermunt, Brand, Bullens, and van Marle (2012) detected different groups of juvenile offenders based on their criminal history.

The LC model was originally known as Latent Structure analysis, but it is also referred to as a binomial (finite) mixture model and it can be seen as the categorical data analogue to factor analysis. Lazarsfeld (1950) introduced it as a method to build a typology or clustering based on a set of dichotomous variables. Much later, Goodman (1974) developed an algorithm to obtain maximum-likelihood estimates and solved identification issues associated with the model, while (Haberman, 1979) showed how the model can be specified as a log-linear model for the contingency table cross-tabulating the categorical latent and observed variables. The algorithm proposed by Goodman, which was later on labelled as the EM algorithm (Dempster, Laird, & Rubin, 1977), is still the dominant approach used for parameter estimation, though it is nowadays often combined with a Newton-Raphson algorithm (Vermunt & Magidson, 2013).

The two basic assumptions of the LC model are that the population consists of K latent classes and that the observed variables are locally independent. The latter implies that responses are assumed to be statistically independent from another within each latent class. The number of classes is usually an unknown "parameter" which is determined by increasing its number as long as some fit measure, e.g. the AIC (Akaike, 1974) or BIC (Schwarz, 1978), improves. The encountered classes are characterized by their class proportions and their response probabilities of all observed variables. Substantive interpretation is given to the latent classes by examining these conditional response probabilities of each class.

With the increasing availability of LC software programs during the 90's, LC analysis became available to the applied researcher and since then the use of LC analysis has increased rapidly, especially in the past two decades. This also led to quite some extensions of the basic LC model, which will be described in the next section.

1.2 Extensions of the LC model

Different types of extensions have been proposed for the basic LC model. Some examples are the LC discrete-factor model which contains multiple categorical latent variables (Magidson & Vermunt, 2001), models with both categorical and continuous latent variables (Dolan & van der Maas, 1998; McLachlan & Peel, 2004; Rost, 1990; Yung, 1997), and multilevel LC models incorporating discrete latent variables at multiple levels of a hierarchical data structure (Nagelkerke, Oberski, & Vermunt, 2016, 2017; Vermunt, 2003). Besides such extensions, LC models have also been used and new LC models have been developed for completely other purposes than substantive interpretation, like density estimation (Van der Palm, van der Ark, & Vermunt, 2016) and multiple imputation (Vidotto, Kaptein, & Vermunt, 2015). Nevertheless, substantive interpretation is still often an important aspect for researchers to choose to perform a LC analysis. However, the results of a LC model are sometimes difficult to interpret. Several alternative extensions have been suggested that facilitate the interpretation of LC models in practical situations.

When the interpretation of the LC models is difficult, it is frequently assessed whether some model assumptions should be relaxed. It is possible to assess what the effect of relaxing a restriction has on the model and its parameters (Oberski, van Kollenburg, & Vermunt, 2013; Oberski, Vermunt,

& Moors, 2015). To accommodate studies with small sample sizes and/or sparse contingency tables, special measures have been proposed to detect misfit (Van Kollenburg, Mulder, & Vermunt, 2015). One of the most common options is to relax the conditional independence assumption between certain pairs of variables (Hagenaars, 1988) by including direct effects in the model. In a confirmatory setting, the number of classes may be based on a priori knowledge, though the specified LC model may not fit due to, for instance, the presence of subclasses or other kinds of mechanisms causing violations of the local independence assumption. Many of these restricted LC models are very similar to (non)parametric IRT models (Croon, 1990; Heinen, 1996; Lindsay, Clogg, & Grego, 1991). However, these options usually require some a priori knowledge and therefore are often not sensible in an exploratory setting.

Furthermore, a LC analysis is frequently only part of the data analysis in a research project, as researchers frequently want to relate the LCs to external variables. It was initially suggested to adapt the standard LC model to include variables affecting the responses (Wedel & DeSarbo, 1994) or the class memberships (Dayton & Macready, 1988). However, this one-step approach is in practice hardly ever used , because applied researchers prefer a separate measurement part (assessing the LC model) and a structural part (relate the LCs to explanatory variables). Therefore the three-step approach is most often used, in which first a LC model is assessed, subsequently class assignment takes place and finally the relation between the LCs and the explanatory variables is assessed. This last step can be corrected for bias caused by classification errors (Bakk, Oberski, & Vermunt, 2016; Bakk & Vermunt, 2016; Bolck, Croon, & Hagenaars, 2004; Vermunt, 2010).

Though various procedures have been developed that help to interpret LC models, there are still situations where interpretation of the LCs can be troublesome. For instance, with large data sets (with a large number of respondents and variables) the fit usually improves until the model contains a large number of classes, as a large number of dependencies needs to be taken into account. This causes many very specific classes to be identified and such specific classes might not be of interest for the research at hand. Moreover, the choice of criterion (e.g., AIC or BIC) can lead to a completely different number of classes. This is even more problematic because different latent class solutions are substantially very hard to compare. This can be seen at the left of Figure 1.1: Every class that is added to a standard LC model constructs a completely new set of classes. Hence, it might very well be that

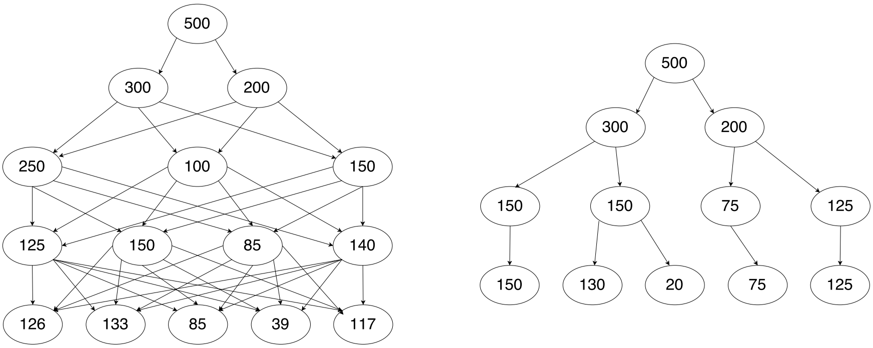

Figure 1.1: Example of a standard LC analysis on the left and a LCT analysis on the right.

none of the K classes is similar to any of the $K - 1$ classes of the previously estimated model.

1.3 Latent Class Trees

To circumvent the issues mentioned above, the possibility to construct classes that are substantially related is introduced in this thesis. When interpretation of a standard LC analysis becomes problematic, we suggest to impose a hierarchical structure on the classes by constructing a Latent Class Tree (LCT). For this purpose, we use the divisive LC algorithm introduced for density estimation by Van der Palm et al. (2016). LCTs are a variant of model-based recursive partitioning, and as such related to decision trees (Breiman, Friedman, Olshen, & Stone, 1984), SEM trees (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013), and divisive cluster analysis (Everitt, Landau, Leese, & Stahl, 2011). Instead of fitting a single model to an observed data set, the data set is partitioned step by step with respect to a LC model with a restriction on the maximum number of classes at a node[*]. Substantially related classes, as shown at the right of Figure 1.1 are constructed by subsequently re-estimating a LC model on each of the "parent" classes. The partitioning procedure is a soft partitioning based on the posterior class membership probabilities and continues as long as for a partitioned class a chosen fit measure prefers a 2-class over a 1-class solution.

Hierarchical tree structures similar to those obtained with a LCT analysis are very practical as clustering procedures because solutions at different levels of a tree allow different granularity to be extracted during the data

[*] A restriction of 2 classes is initially used, but extensions are discussed in Chapter 3.

analysis, making them ideal for exploration (Ghattas, Michel, & Boyer, 2017; Zhao, Karypis, & Fayyad, 2005). A step-wise interpretation is often easier but, moreover, the interpretation of a split can be used to determine the appropriate number of classes for the topic at hand. For instance, at the bottom of the LCT in Figure 1.1, the choice could be made to disregard the final split based on interpretation of this split. Therefore the LCT methodology can be a practical alternative for LC models when one encounters difficulties in deciding about the number of classes or in interpreting the differences between a large number of classes.

1.4 Outline of the dissertation

This thesis consists of four papers in which different aspects of the new LCT approach are developed. In each of these papers, the proposed LCT models are illustrated by empirical examples, which is the most practical way to show the benefits for interpretation and substantive assessment of the number of classes of a LCT.

- The second chapter of this dissertation contains an introduction to LCTs. Based on the divisive LC algorithm for density estimation by Van der Palm et al. (2016), the basic procedure to build a LCT is described. This basic LCT consisting of binary splits only is illustrated by an empirical example in which a LCT on social capital is built.

- In the third chapter of this dissertation, some problems associated with having only binary splits are discussed. It is discussed how one can decide to increase the number of classes of a split and whether this is the same for the first and subsequent splits of an LCT. Subsequently, a relative measure of fit is introduced to decide whether to increase the number of classes of the first split of an LCT. This approach is illustrated again with the empirical example on social capital. Moreover, the LCT procedure is also applied to data from a cross-national study using a set of ranking items on (post-)materialism, which illustrates how it can be accommodated for other types of LC models.

- The fourth chapter of this dissertation extends the LCT procedure to the longitudinal framework. By applying the LCT procedure in the context of Latent Class Growth modeling it is shown how a Latent Class Growth Tree (LCGT) can be constructed. For longitudinal data the tree approach is even more useful, because such applications result even more often in a large number of classes than LC models based on cross sectional data. However, it also requires some additional considerations, such as the specification of the shape of the trajectories. These LCGT models are illustrated by empirical examples on drugs use during adolescence and mood regulation during the day assessed using experience sampling.

- The fifth and final chapter of this dissertation extends the use of LCTs by showing what can be done after building a tree and how the classes of a tree can be related to covariates. It is shown how distal outcomes are related to the classes and how class membership can be predicted based on covariates. Both options are illustrated with empirical examples, one on social capital and one on mood regulation.

The four chapters of this thesis were written as separate articles intended for publication in academic journals. Because the content of each chapter was kept as close to the original articles, the chapters contain some overlap.

Chapter 2

Building Latent Class Trees, with an application to a study of social capital

Abstract

Researchers use latent class analysis to derive meaningful clusters from sets of categorical observed variables. However, especially when the number of classes required to obtain a good fit is large, interpretation of the latent classes in the selected model may not be straightforward. To overcome this problem, we propose an alternative way of performing a latent class analysis, which we refer to as latent class tree modelling. For this purpose, we use a recursive partitioning procedure similar to those used in divisive hierarchical cluster analysis; that is, classes are split until the model selection criterion indicates that the fit does no longer improve. The key advantage of the proposed latent class tree approach compared to the standard latent class analysis approach is that it gives a clear insight into how the latent classes are formed and how solutions with different numbers of classes are linked to one another. We also propose measures to evaluate the relative importance of the splits. The practical use of the new approach is illustrated by the reanalysis of a data set with indicators of social capital.

This chapter is published as Bergh, M. van den, Schmittmann, V.D., & Vermunt, J.K. (2017) Building latent class trees, with an application to social capital *Methodology*, *12*(4), 124-138

2.1 Introduction

Latent class (LC) analysis has become a popular statistical tool for identifying subgroups or clusters of respondents using sets of observed categorical variables (Clogg, 1995; Goodman, 1974; Hagenaars, 1990; Lazarsfeld & Henry, 1968; McCutcheon, 1987). Since in most LC analysis applications the number of subgroups is unknown, the method will typically be used in an exploratory manner; that is, a researcher will estimate models with different numbers of latent classes and select the model which performs best according to a certain likelihood-based criterion, for instance, the BIC or AIC. Although there is nothing wrong with such a procedure, in practice it is often perceived as being problematic, especially when the model is applied to a large data set; that is, when the number of variables and/or the number of subjects is large. One problem occurring in such situations is that the selected number of classes may be rather large, which makes their interpretation difficult. A second problem results from the fact that usually one would select a different number of classes depending on the model selection criterion used, and that because of this, one may wish to inspect multiple solutions because each of them may reveal specific relevant features in the data. However, it is often unclear how solutions with different numbers of classes are connected, making it very hard to see what a model with more classes adds to a model with less classes.

To overcome the above mentioned problems, we propose an alternative way of performing a latent class analysis, which we call Latent Class Tree (LCT) modeling. More specifically, we have developed an approach in which a hierarchical structure is imposed on the latent classes. This is similar to what is done in hierarchical cluster analysis (Everitt et al., 2011), in which clusters are either formed by merging (the agglomerative procedure) or splitting (the divisive procedure) clusters which were formed earlier. For hierarchical cluster analysis it has been shown that divisive procedures work at least as well as the more common agglomerative procedures in terms of both computational complexity and cluster quality (Ding & He, 2002; Zhao et al., 2005). Here, we will use a divisive procedure in which latent classes are split step-by-step since such an approach fits better with the way LC models are estimated than an agglomerative approach.

For the construction of a LCT we use the divisive LC analysis algorithm developed by Van der Palm et al. (2016) for density estimation, with applications in among others missing data imputation. This algorithm starts with a parent node consisting of the whole data and involves estimating a 1- and a 2-class model for the subsample at each node of the tree. If a 2-class model is preferred according to the fit measure used, the subsample at the node concerned is split and two new nodes are created. The procedure is repeated at the next level of the hierarchical structure until no further splits need to be performed. Van der Palm et al. (2016) used this algorithm with the aim to estimate LC models with many classes, say 100 or more, in an efficient manner. Because they were not interested in the interpretation of the classes but only in obtaining an as good as possible representation the data, they used very liberal fit measures. In contrast, our LCT approach aims at yielding an interpretable set of latent classes. In order to construct a substantively meaningful and parsimonious tree, we will use the rather conservative BIC (Schwarz, 1978) to decide about a possible split.

The resulting tree structure contains classes which are substantively linked. Pairs of lower-order classes stem from a split of a higher-order class and vice versa a higher-order class is a merger of a pair of lower-order classes. The tree structure can be interpreted at different levels, where the classes at a lower level yield a more refined description of the data than the classes at a higher level of the tree. To further facilitate the interpretation of the classes at different levels of the tree, we have developed a graphical representation of the LCT, as well as propose measures quantifying the relative importance of the splits. It should be noted that the proposed LCT approach resembles the well-known classification trees (Friedman, Hastie, & Tibshirani, 2001; Loh & Shih, 1997) in which at each node it is decided whether the subsample concerned should be split further. Classification trees are supervised classification tools in which the sample is split based on the best prediction of a single outcome using a set of observed predictors variables. In contrast, the LCT is an unsupervised classification tool, in which the sample is split based on the associations between multiple response variables rather than on observed predictors.

Two somewhat related approaches for imposing a hierarchical structure on latent classes have been proposed before. Zhang (2004) developed a hierarchical latent class model aimed at splitting the observed variables into sets, where each set is linked to a different dichotomous latent variable and where the dependencies between the dichotomous latent variables are modeled by a tree structure. The proposed LCT model differs from this approach in that it aims at clustering respondents instead of variables. Hennig (2010) proposed various methods for merging latent classes derived from a set of continuous variables. His approach differs from ours in that it uses an agglomerative instead of a divisive approach and, moreover, that it requires applying a standard latent class model to select a solution from which the merging should start. Though LCT modeling may also be applicable with continuous variables, here we will restrict ourselves to its application with categorical data.

The next section describes the algorithm used for the construction of a LCT in more detail and presents post hoc criteria to evaluate the importance of each split. Subsequently, the use of the LCT model is illustrated using an application to a large data set with indicators on social capital. A discussion on the proposed LCT method is provided in the last section.

2.2 Method

2.2.1 Standard LC analysis

Let y_{ij} denote the response of individual i on the jth of J categorical response variables. The complete vector of responses of individual i is denoted by $\mathbf{y}_i$. A latent class analysis defines a model for the probability of observing $\mathbf{y}_i$; that is, for $P(\mathbf{y}_i)$. Denoting the discrete latent class variable by X, a particular latent class by k, and the number of latent classes by K, the following model is specified for $P(\mathbf{y}_i)$:

$$P(\mathbf{y}_i) = \sum_{k=1}^{K} P(X = k) \prod_{j=1}^{J} P(y_{ij}|X = k). \tag{2.1}$$

Here, $P(X = k)$ represents the (unconditional) probability of belonging to class k and $P(y_{ij}|X = k)$ represents the probability of giving the response concerned conditional on belonging to class k. The product over the class-specific response probabilities shows the key model assumption of local independence.

LC models are typically estimated by maximum likelihood, which involves finding the values of the unknown parameters maximizing the following log-likelihood function:

$$\log L(\theta; \boldsymbol{y}) = \sum_{i=1}^{N} \log P(\mathbf{y}_i), \tag{2.2}$$

where N denotes the total sample size and where $P(\mathbf{y}_i)$ takes the form defined in Equation (2.1). Maximization is typically done by means of the EM algorithm.

2.2.2 Building a LCT

The building of a LCT involves the estimation and comparison of 1- and 2-class models only. If a 2-class solution is preferred over a 1-class solution (say based on the BIC), the sample is split into two subsamples and 1- and 2-class models will subsequently be estimated for both newly formed samples. This top-down approach continues until only 1-class models are preferred, yielding the final hierarchically ordered LCT. An example of such a LCT is depicted in Figure 2.1. The top level contains the root node which consists of the complete sample. After estimating 1- and 2-class models with the complete sample, it is decided that the 2-class model is preferred, which implies that the sample is split into two subsamples (class X=1 and class X=2), which form level 2 of the tree. Subsequently, class 1 is split further while class 2 is not, yielding classes X_1=1, X_1=2, and X_2=1 at level 2. In our example, after level 4 there are no splits anymore and hence the final solution can be seen at both levels 4 and 5. Though level 5 is redundant, this is only visible after the procedure has been finished; i.e., after only 1-class models are preferred.

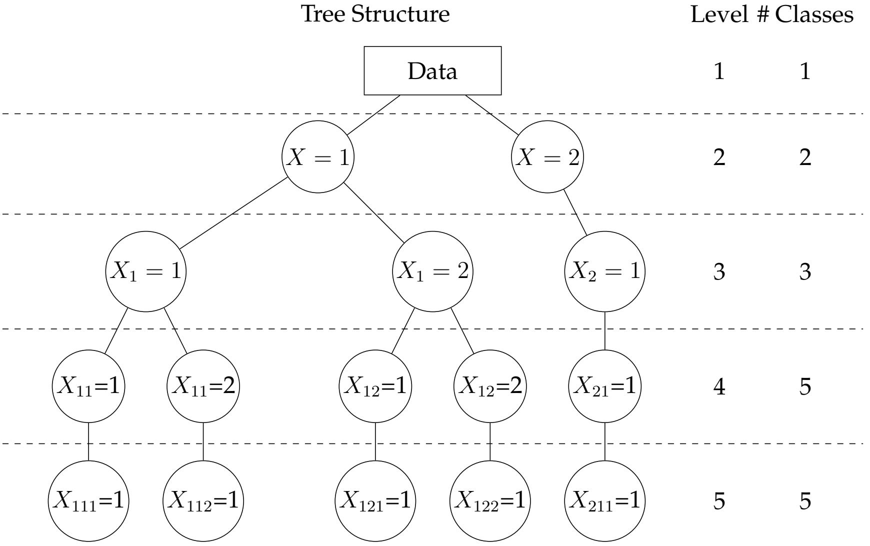

Figure 2.1: Graphical example of a LCT

More formally, the 2-class LC model defined at a particular parent node can be formulated as follows:

$$P(\mathbf{y}_i|X_{parent}) = \sum_{k=1}^{2} P(X_{child} = k|X_{parent}) \prod_{j=1}^{J} P(y_{ij}|X_{child} = k, X_{parent}) \quad (2.3)$$

where X_{parent} represent the parent class at level t and X_{child} one of the two possible newly formed classes at level $t+1$. In other words, as in a standard LC model we define a model for $\mathbf{y}_i$, but now conditioning on belonging to the parent class concerned.

A key issue for the implementation of the divisive LC algorithm illustrated in Figure 2.1 is how to perform the split at the parent node when a 2-class model is preferred. As proposed by Van der Palm et al. (2016), we use a proportional split based on the posterior class membership probabilities for the two child nodes conditional on the parent node, denoted by $k = 1, 2$. These are obtained as follows:

$$P(X_{child} = k|\mathbf{y}_i; X_{parent}) = \frac{P(X_{child} = k|X_{parent}) \prod_{j=1}^{J} P(y_{ij}|X_{child} = k, X_{parent})}{P(\mathbf{y}_i|X_{parent})} \quad (2.4)$$

Estimation of the LC model at the parent node X_{parent} involves maximizing the following weighted log-likelihood function:

$$\log L(\theta; \boldsymbol{y}, X_{parent}) = \sum_{i=1}^{N} w_{i,X_{parent}} P(\mathbf{y}_i|X_{parent}) \quad (2.5)$$

where $w_{i,X_{parent}}$ is the weight for person i at the parent class, which equals the posterior probability of belonging to the parent class for the individual concerned. If a split is performed, the weights for the two newly formed classes at the next level are obtained as follows:

$$w_{i,X_{child}=1} = w_{i,X_{parent}} P(X_{child} = 1|\mathbf{y}_i; X_{parent}) \quad (2.6)$$

$$w_{i,X_{child}=2} = w_{i,X_{parent}} P(X_{child} = 2|\mathbf{y}_i; X_{parent}). \quad (2.7)$$

In other words, a weight at a particular node equals the weight at the parent node times the posterior probability of belonging to the child node concerned conditional on belonging to the parent node. As an example, the weights

$w_{i,X_1=2}$ used for investigating a possible split of class $X_1 = 2$ are constructed as follows:

$$w_{i,X_{12}} = w_{i,X=1} P(X_1 = 2|\mathbf{y}_i, X = 1), \tag{2.8}$$

where in turn $w_{i,X=1} = P(X = 1|\mathbf{y}_i)$. This implies:

$$w_{i,X_{12}} = P(X = 1|\mathbf{y}_i) P(X_1 = 2|\mathbf{y}_i, X = 1), \tag{2.9}$$

which shows that a weight at level 2 is in fact a product of two posterior probabilities.

Construction of a LCT can thus be performed using standard software for LC analysis, namely by running 1- and 2-class models multiple times with the appropriate weights. We developed an R routine in which this process is fully automated[*]. It calls the Latent GOLD program (Vermunt & Magidson, 2013) in batch mode to estimate the 1- and 2-class models, evaluates whether a split should be made, and keeps track of the weights when a split is accepted. In addition, it creates several types of graphical displays which facilitate the interpretation of the LCT. A very useful and novel graphical display is a tree depicting the class-specific response probabilities $P(y_{ij}|X_{child} = k, X_{parent})$ for the newly formed child classes using profile plots (for an example, see Figure 2.2). In this tree, the name of a child class equals the name of the parent class plus an additional digit, a 1 or a 2. To prevent that the structure of the tree will be affected by label switching resulting from the fact the order of the newly formed classes depends on the random starting values, when building the LCT we locate the larger class at the left branch with number 1 and the smaller class at the right branch with number 2.

2.2.3 Statistics for building and evaluating a LCT

In a standard LC analysis, one will typically estimate the model for a range of values for K, say from 1 to 10, and select the model that performs best according to the chosen fit measure. The most popular measures are information criteria such as BIC, AIC, and AIC3, which aim at balancing model fit and parsimony (Andrews & Currim, 2003; Nylund, Asparouhov, & Muthén, 2007). Denoting the number of parameters by P, these measures are defined as follows:

[*] Though still under development, this can be retrieved from `http://github.com/MattisvdBergh/LCT`

$$\begin{aligned} BIC &= -2\log L + \log(N)P && (2.10) \\ AIC &= -2\log L + 2P && (2.11) \\ AIC3 &= -2\log L + 3P && (2.12) \end{aligned}$$

Because $\log(N)$ is typically larger than 3, the BIC penalizes the number of parameters most strongly. This implies that BIC will select a model with a smaller than or equal number of classes as AIC3, and AIC3 with a smaller than or equal number of classes as AIC.

As in a standard LC model, at each parent node that can potentially be split, we need to determine which model should be preferred, with the difference that here we only have to make a choice between a 1- and a 2-class model. In the empirical example presented in the next section, we will base this decision on the BIC, which means that we give a large weight to parsimony. However, in the evaluation of the tree, we will also investigate which splits rejected by BIC would be accepted by AIC3. In the computation of the BIC, we use the total sample size, and thus not the sample size at the node concerned. Note that classes are split as long as the difference between the BIC of the estimated 1- and 2-class models, $\Delta BIC = BIC(1) - BIC(2)$, is larger than 0. The size of ΔBIC can be compared across splits, where larger ΔBIC values indicate that a split is more important; that is, it yields a larger increase of the log-likelihood and thus a larger improvement of fit.

Another possible way to assess the importance of a split is by looking at the reduction of a goodness-of-fit measure such as the Pearson chi-square. Because overall goodness-of-fit measures are not very useful when the number of response variables is large, we will use a measure based of the fit in two-way tables. The fit in a two-way table can be quantified using the bivariate residual (BVR), which is a Pearson chi-square statistic divided by the number of degrees of freedom (Oberski et al., 2013). A large BVR value indicates that the association between that pair of variables is not picked up well by the LC model or, alternatively, that the local independence assumption does not hold for the pair concerned. By summing the BVR values across all pairs of variables, we obtain what Van Kollenburg et al. (2015) refer to as the total BVR (TBVR):

$$TBVR = \sum_{j=1}^{J} \sum_{j'=1}^{j-1} BVR_{jj'} \tag{2.13}$$

A split is more important if it yields as larger reduction of the $TBVR$ between the 1- and 2-class solution. In other words, we look at: $\Delta TBVR = TBVR(1) - TBVR(2)$.

While ΔBIC and $\Delta TBVR$ can be used to determine the importance of the splits in terms of model fit, it may also be relevant to evaluate the quality of splits in terms of their certainty or, equivalently, in terms of the amount of separation between the child classes. This is especially relevant if one would like to assign individuals to the classes resulting from a LCT. Note that the assignment of individuals to the two child classes is more certain when the larger of the posterior probabilities $P(X_{child} = k|\mathbf{y}_i; X_{parent})$ is closer to 1. A measure to express this is the entropy; that is,

$$Entropy(X_{child}|\mathbf{y}) = \tag{2.14}$$
$$\sum_{i=1}^{N} w_{i|X_{parent}} \sum_{k=1}^{2} -P(X_{child} = k|\mathbf{y}_i; X_{parent}) \log P(X_{child} = k|\mathbf{y}_i; X_{parent}).$$

Typically $Entropy(X_{child}|\mathbf{y})$ is rescaled to lie between 0 and 1 by expressing it in terms of the reduction compared to $Entropy(X_{child})$, which is the entropy computed using the unconditional class membership probabilities $P(X_{child} = k|X_{parent})$. This so-called $R^2_{Entropy}$ is obtained as follows:

$$R^2_{Entropy} = \frac{Entropy(X_{child}) - Entropy(X_{child}|\mathbf{y})}{Entropy(X_{child})} \tag{2.15}$$

The closer $R^2_{Entropy}$ is to one, the better the separation between the child classes in the split concerned.

2.3 Application of a LCT to a study of social capital

2.3.1 Building the LCT

The proposed LCT methodology is illustrated by a reanalysis of a large data set which was previously analyzed using a standard LC model. Owen and Videras (2008) used the information from 14.527 respondents of the 1975, 1978, 1980, 1983, 1984, 1986, 1987 through 1991, 1993, and 1994 samples of the General Social Survey to construct “a typology of social capital that accounts for the different incentives that networks provide.” The data set contains sixteen dichotomous variables indicating whether respondents participate in specific types of voluntary organizations (the organizations are listed in the legend of Figure 2.2) and two variables indicating whether respondents agree with the statements "other people are fair" and "other people can be trusted". Owen and Videras explain the inclusion of the latter two variables by stating that social capital is a multidimensional concept which embeds multiple manifestations of civic engagement as well as trust and fairness. Using the BIC, Owen and Videras selected a model with eight classes, while allowing for one local dependency, namely between fraternity and school fraternity.

Figure 2.2 depicts the results obtained when applying our LCT approach using the BIC as the splitting criterion. A figure of a tree containing information on the sample sizes a the different nodes is provided in Appendix A.1. As can be seen, at the first two levels of the tree, all classes are repetitively split. However, at the third level only three out of four classes are split, as a division of class 12 is not supported by the BIC. Subsequently, the number of splits decreases to two at the fourth level, while at the fifth level there are no more splits, indicating the end of the divisive procedure.

For the interpretation of the LCT, we can use the profile plots, which show which variables are most important for the split concerned (exact probabilities can be found in Appendix A.1). From the upper panel of Figure 2.2, which depicts class-specific response probabilities for classes 1 and 2, it can easily be seen that all probabilities are higher for class 2 than for class 1, which is confirmed by Wald tests ($W \geq 7.43$, $p < 0.05$). So basically the first split divides the sample based on general social capital, where class 1 contains respondents with low social capital and class 2 respondents with high social capital. This is supported by the total group participation of each class (TGP, the sum of all probabilities except fair and trust), which equals 0.88 for class 1 and 3.83 for class 2.

The second row of Figure 2.2 shows the splitting of both class 1 and 2 is mainly due to the variables fair and trust. Apparently the low and high social capital groups can both be split based on how respondents view other people regarding fairness and trustworthiness. This categorization will be called optimists versus pessimists. The difference in TGP is relatively small for these two splits, being 0.09 between class 11 and 12 and 0.83 between class 21 and 22. Up to here, there are four classes: pessimists with low social capital (11), optimists with low social capital (12), optimists with high social capital (21) and pessimists with high social capital (22).

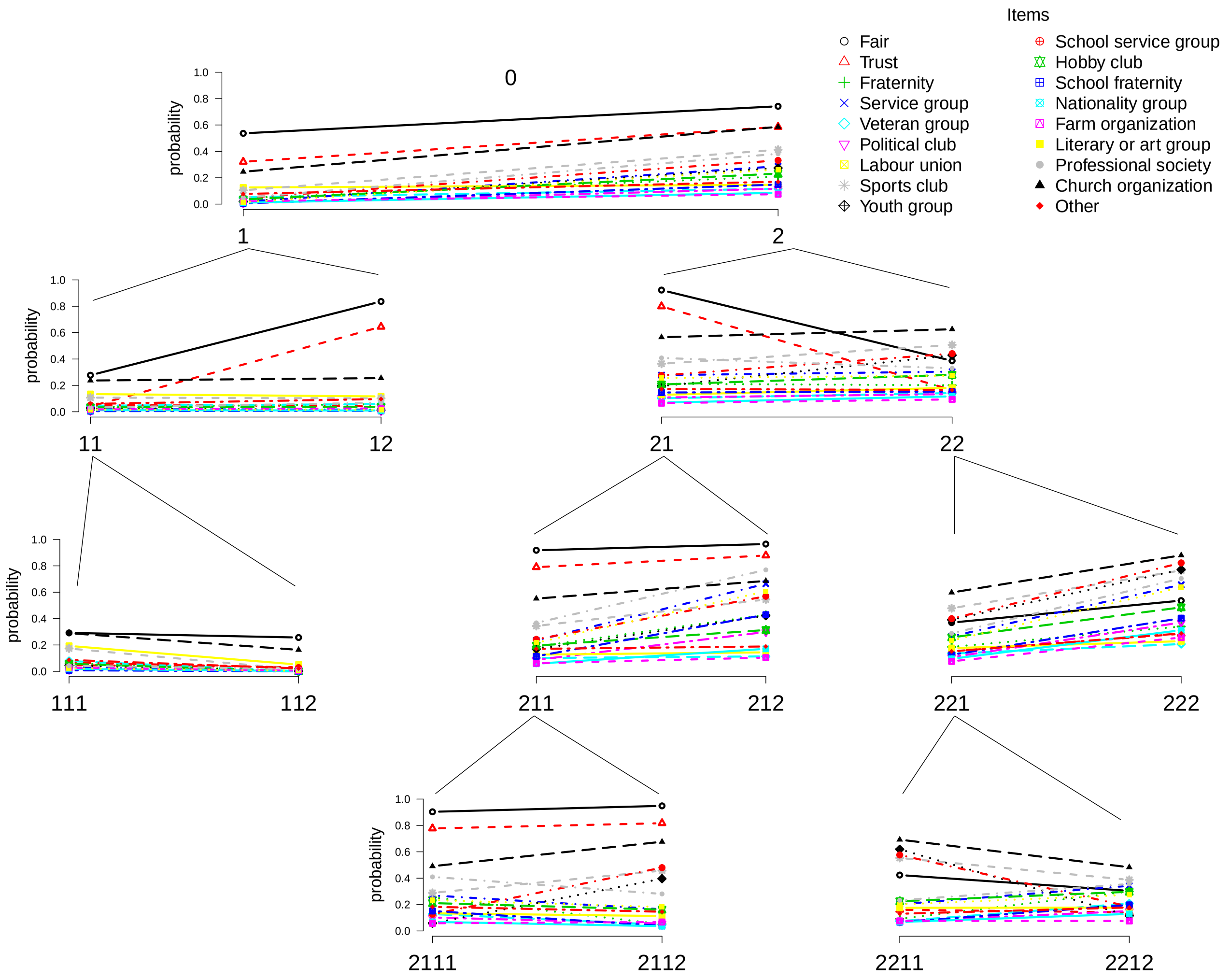

Figure 2.2: Profile plots of LCT on social capital

Looking at the next level, one can see that class 12 is not split further. The third row of Figure 2.2 shows similar patterns for all three splits at this level: all probabilities are lower in one class than in the other. Therefore these splits can be interpreted as capturing more refined quantitative differences in social capital. This results in seven classes, ranging from high to very low social capital, as can be seen from the TGP values reported in Table 2.1.

Table 2.1: Interpretation of classes at level 3 with TGP in brackets

Social capital	Pessimist	Optimist
High	222 (8.13)	212 (6.45)
Average	221 (3.97)	211 (3.23)
Low	111 (1.22)	12 (0.93)
Very Low	112 (0.31)	

At the fourth level, both the optimists and pessimists class with average social capital (211 & 221) are split. Contrary to the previous splits, here we can see qualitative differences in terms of the type of organization in which one participates. For instance, in classes 2112 and 2211, respondents have higher probabilities of being a member of a sports or a youth group, while in the corresponding classes 2111 and 2212, respondents have a higher probability of being a member of a professional organization. The TGP of the newly formed classes ranges from 3.17 to 4.06, while fair and trust are high at the optimistic branch and low at the pessimistic branch of the tree. At level five no further splits occur.

At the lowest level, the constructed LCT has nine classes, one more than obtained with a standard LC analysis. It turns out that the classes identified with the two alternative approaches are rather similar. The parameters from the standard 8-class model appear in the profile plot depicted in Figure 2.3 and in Appendix A.2. For instance, the conditional probabilities of LC-class 1 are very similar to those of LCT-classes 111 and 112. Moreover, LC-class 1 is even more similar to the higher-order LCT-class 11, which suggests that the distinction between LCT-classes 111 and 112 is probably not made in the standard LC analysis. The three largest classes of the original analysis are very similar to at least one LCT-class (LC 1 to TLC 11, LC 2 to LCT 12 and LC 3 to LCT 2111), while 3 out of the five smaller original classes can also be directly related to a LCT-class (LC 6 to LCT 221, LC 7 to LCT 2112 and LC 8 to LCT 222). LC-classes 4 and 5 (containing 7% and 5% of the respondents) are not clearly related to a LCT-class.

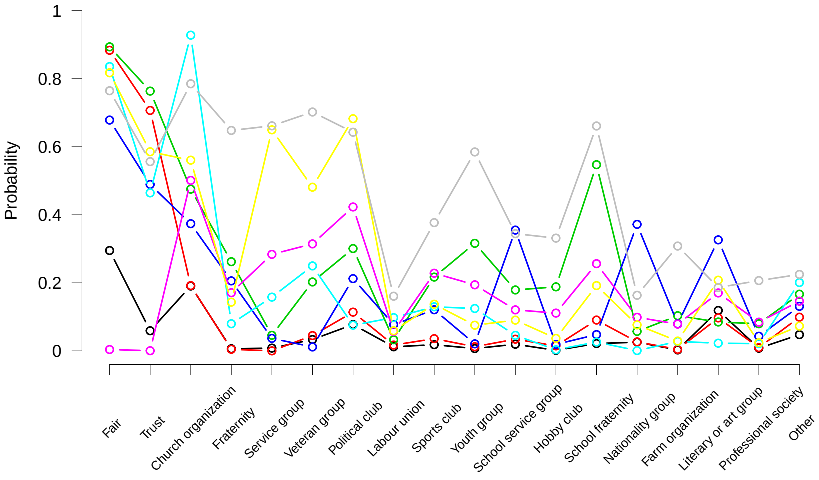

Figure 2.3: Profile plot of original LC solution

2.3.2 Evaluating the splits of the LCT

Now let us look in more detail at the model fit and classification statistics associated with the accepted and rejected splits. Table 2.2 reports the values of ΔBIC, $\Delta AIC3$, $\Delta TBVR$, and $R^2_{Entropy}$, as well as the class proportions, for the considered splits, where the classes split based on the ΔBIC appear in the top rows and the others in the bottom rows. Looking at the $\Delta AIC3$, we can see that this criterion would have allowed (at least) five additional splits. The $\Delta TBVR$ values show the fit always improves, but the improvements are larger for the accepted than for the rejected splits. The $R^2_{Entropy}$ indicating the quality of a split in terms of classification performance, shows a rather different pattern: it takes on both higher and lower values among accepted and non-accepted splits.

Based on the information provided in Table 2.2, one could opt not to split class 11. Compared to other accepted splits, splitting this class contributes much less in terms of improvement of fit, while also the classification performance associated with this split is rather bad. Note also that this is one of the largest classes and therefore the statistical power to retrieve subclasses with small differences is relatively high. The decision on retaining this split depends on the whether the encountered more detailed distinction within this low social capital and pessimistic class is of substantive interest. However, what is clear is that if a good classification performance is required, this split seems to be less appropriate.

Table 2.2: Information criteria per split, with split classes in the top and not split classes in the bottom rows

	ΔBIC	$\Delta AIC3$	$\Delta TBVR$	$R^2_{Entropy}$	$P(X=k)$
0	9205.4	9330.5	23597.7	0.648	1.000
1	1346.2	1471.3	1495.0	0.489	0.705
2	691.0	816.1	1071.4	0.516	0.295
11	30.8	155.9	279.1	0.261	0.378
21	117.7	242.8	275.6	0.512	0.195
22	94.9	220.0	285.2	0.610	0.100
211	92.9	218.0	338.2	0.353	0.176
221	58.6	183.7	313.9	0.433	0.090
12	-37.7	87.4	179.4	0.222	0.327
111	-84.3	40.8	100.5	0.295	0.221
112	-167.3	-42.2	16.4	0.174	0.157
212	-125.5	-0.4	72.3	0.473	0.020
222	-119.0	6.1	64.6	0.815	0.010
2111	-2.7	122.4	206.0	0.353	0.118
2112	-126.7	-1.6	63.7	0.288	0.058
2211	-136.4	-11.4	54.8	0.257	0.049
2212	-99.1	26.0	100.6	0.383	0.041

Conversely, one might want to include the split of class 2111. Though this split was rejected by the ΔBIC stop criterion, this is based on a rather small negative value, while the values for the $\Delta AIC3$ and $\Delta TBVR$ are relatively high. However, the $R^2_{Entropy}$ indicates a low quality of this split. Hence, the information on the fit improvement might be misleading, due to this class being the largest class at the lowest level of the tree.

The opposite is true for the split of class 222. Though this class is quite small and the fit statistics of this split indicate not much improvement, the $R^2_{Entropy}$ indicates that classes 2221 and 2222 would be very well separated. Of course, once again the research question at hand is crucial for the decision to add a class to the tree. For exploration the split of class 2111 can be relevant, while for classification the split of class 222 might be more appropriate.

2.4 Discussion

In this paper, we proposed an alternative way of performing a latent class analysis, which we called Latent Class Tree modeling. More specifically, we showed how to impose a hierarchical structure on the latent classes using the divisive LC analysis algorithm developed by Van der Palm et al. (2016). To further facilitate the interpretation of the classes created at different levels of the tree, we developed graphical representations of the constructed LCT, as well as proposed measures quantifying the relative importance and the quality of the splits. The usefulness of the new approach was illustrated by an empirical example on latent classes differing in social capital using data from the General Social Survey.

Various issues related to the construction of LCTs need further study. The first we would like to mention is related to the fact that we choose to restrict ourselves to binary splits. However, the LCT can easily be extended to allow for splits consisting of more than two classes. It is not so difficult to think of situations in which it may be better to start with a split into say three or four classes, and subsequently continue with binary splits to fine tune the solution. The main problem to be resolved is what kind of statistical criterion to use for deciding about the number of classes needed at a particular split. One cannot simply use the BIC, since that would again yield a standard LC model.

In the empirical application, we used the BIC based on the total sample size as the criterion for deciding whether a class should be split. However, the use of a more liberal criterion may make sense in situations in which the research question at hand requires more detailed classes. Criteria such as the AIC3 or the BIC based on the sample size at the node concerned will result in a larger and more detailed tree, but the estimates for the higher-order classes will remain the same. At the same time, the stopping criterion for the LCT approach could be made more strict by including additional requirements, such as the minimal size of the parent class and/or the child classes, the minimal classification performance in terms of $R^2_{Entropy}$, or the minimal number of variables providing a significant contribution to a split. The possible improvement of the stopping criterion is another topic that needs further research.

In the current paper, we restricted ourselves to LC models for categorical variables. However, LC models have also become popular cluster analysis tools for continuous and mixed response variables (Hennig & Liao, 2013; Vermunt & Magidson, 2002). In these kinds of applications, the number of latent classes obtained using a standard LC analysis can sometimes be rather large. It would therefore be of interest to extend the proposed LCT approach to be applicable in those situations as well.

Chapter 3

Deciding on the starting number of classes of a Latent Class Tree

Abstract

Recently, Latent Class Tree (LCT) modelling has been proposed as a convenient alternative to standard latent class (LC) analysis. Instead of using an estimation method in which all classes are formed simultaneously given the specified number of classes, in LCT analysis a hierarchical structure of mutually linked classes is obtained by sequentially splitting classes into two subclasses. The resulting tree structure gives a clear insight into how the classes are formed and how solutions with different numbers of classes are substantively linked to one another. A limitation of the current LCT modelling approach is that it allows only for binary splits, which in certain situations may be too restrictive. Especially at the root node of the tree, where an initial set of classes is created based on the most dominant associations present in the data, it may make sense to use a model with more than two classes. In this paper, we propose a modification of the LCT approach which allows for a non-binary split at the root node, and provide methods to determine the appropriate number of classes in this first split, either based on theoretical grounds or based on a relative improvement of fit measure. Furthermore, we show how to apply a LCT model when a non-standard LC model is required. These new approaches are illustrated using two empirical applications: one on social capital and another on (post-)materialism.

This chapter is conditionally accepted for publication in Sociological Methodology as Bergh, M. van den, Kollenburg, G.H. van, & Vermunt, J.K. Deciding on the starting number of classes of a Latent Class Tree

3.1 Introduction

Latent Class (LC) modelling has become a popular tool for clustering respondents into homogeneous subgroups based on their responses on a set of categorical variables (Clogg, 1995; Goodman, 1974; Hagenaars, 1990; Lazarsfeld & Henry, 1968; Magidson & Vermunt, 2004; McCutcheon, 1987; Vermunt & Magidson, 2002). LC models have been applied for the investigation of a variety of subjects, e.g., risk behavior like gambling (Studer et al., 2016) and suicide attempts (Thullen, Taliaferro, & Muehlenkamp, 2016), social constructs like social class (Savage et al., 2013) and social support (Santos, Amorim, Santos, & Barreto, 2015), and cognitive constructs like rule assessment (Jansen & van der Maas, 1997) and cognitive control (Van Hulst, de Zeeuw, & Durston, 2015).

A crucial part of doing a LC analysis is the decision on the required number of classes. In a confirmatory setting, the number of classes may be based on a priori knowledge, though the specified LC model may not fit due to, for instance, the presence of subclasses or other kinds of mechanisms causing violations of the local independence assumption. In such situations, it may make sense to relax the local independence assumption, as suggested among other by Oberski (2016).

In an exploratory setting, we will typically not aim at finding the "true" number of clusters, but instead look for a clustering that describes the data reasonably well and is moreover easy to interpret. To achieve this goal, researchers estimate models with different numbers of classes and select the model that performs best according to some fit measure, for example, according to the information criterion AIC or BIC. While AIC and BIC penalize model complexity and thus prefer models with less classes, when applying LC models to data sets which are (very) large in terms of number of cases and/or number of variables, one will often end up with a model with a large number of classes. Some of these classes may differ from one another in very specific and possibly less interesting ways, making their distinction hard to interpret substantively. Moreover, different model selection measures will typically point at different best models in terms of the number of classes. In such situations, researchers can no longer rely on purely statistical criteria, but will instead need to inspect solutions with different number of classes and probably opt for the model that fits best to their substantive goals (e.g., Hadiwijaya, Klimstra, Vermunt, Branje, & Meeus, 2015; Oser, Hooghe,

& Marien, 2013; Spycher, Silverman, Brooke, Minder, & Kuehni, 2008; Sullivan, Kessler, & Kendler, 1998). It will be clear that such an approach may be somewhat problematic since different researchers may come up with rather different final models when analyzing exactly the same data.

To overcome the abovementioned problems associated with LC analysis applications with large data set sets, Van den Bergh, Schmittmann, and Vermunt (2017) proposed an alternative way of performing a LC analysis, which they called LC Tree (LCT) analysis. Their approach involves performing a divisive hierarchical cluster analysis using an algorithm develop by Van der Palm et al. (2016) for density estimation with a large number of categorical variables. The main advantage of the LCT modelling approach is that it shows how models with different numbers of classes are linked to one another; for instance, a model with 6 classes is a model with 5 classes in which one of the classes split into two parts. When applying a LCT, the model selection problem reduces to deciding whether a particular split should be accepted yes or no. As in a standard LC analysis, this can be decided based on fit measures, but also based on whether a split is meaningful content wise.

As the name suggests, the method yields tree structure (see Figure 3.1 for an example), which at the top contains a root node that serve as ’parent’ node of two ’child’ nodes. At the next level of the tree, these child nodes become parent nodes and produce possibly their own child nodes, and so on. More specifically, the algorithm used to construct a LCT works as follows: first a 1- and 2-class model is estimated for the root node, that is, using the original data set. If the 2-class model is preferred according to the model selection criterion used, then two child nodes are created. For each of the two child nodes a new data set is constructed, which contains the posterior membership probabilities for the class concerned as case weight. Subsequently, each new child node is treated as a parent and it is checked whether a 2-class model provides a better fit than a 1-class model on the corresponding weighted data set. This stepwise procedure continues until no node is split up anymore.

The sequential LCT algorithm yields child classes which are subclasses of a parent class, which implies that interpretation can take place at any level of the tree. That is, after labeling the classes formed at the root of the tree, the classes formed at the next level of the tree will be labelled conditionally on the labeling of their parent classes. This makes it much easier to interpret LC solution with more than a few classes. Moreover, the fact the classes are hierarchically linked makes it possible to decide on the number of classes

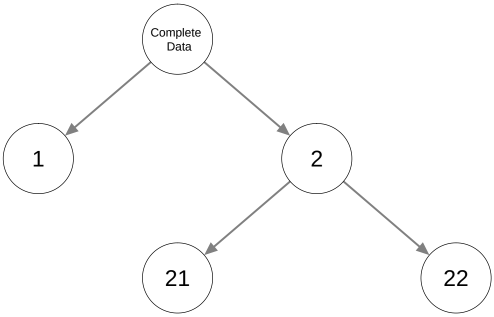

Figure 3.1: Example of a tree structure with two binary splits.

based on substantive interpretation of the splits; if certain splits are not interesting or relevant for the research question at hand, the child classes of a split can be substituted for their parent class. Hierarchical tree structures similar to those obtained with a LCT analysis are very practical as clustering procedures because clustering solutions at different levels of a tree allow different granularity to be extracted during the data analysis, making them ideal for exploration (Ghattas et al., 2017; Zhao et al., 2005).

An important limitation of the current LCT modeling approach is that it is limited to binary splits. While this may be less of a problem for the lower levels of the tree where more detailed between-cluster differences are detected, it can sometimes be problematic for the root of the tree. As an illustration of this problem, Figure 3.2 presents three examples of possible latent class configurations: two with three classes and one with four classes. The first configuration of three classes (Panel A) shows two fairly similar classes (classes 2 and 3), while class 1 is quite distinct from these two. This is a situation in which a tree with binary splits is expected to perform well. In the first binary split, class 1 will be separated from classes 2 and 3, where the class combining the latter two will have response probabilities close to 0.2 (the average of these two classes). The binary split at the next level will detect the differences between class 2 and 3. Hence, binary splits do not cause any problems with this setup and an example of the resulting tree structure is shown by Figure 3.1, where classes 2 and 3 are defined as 21 and 22 in the tree structure.

The second configuration of three classes in Figure 3.2 (Panel B) shows three rather distinct classes. The first binary split will mainly be based on most dissimilar classes 1 and 3, while class 2 will be spread out over the two classes. By splitting both classes again, a third and fourth class are retrieved and a tree structure as shown in Figure 3.3 is obtained. Neither the number

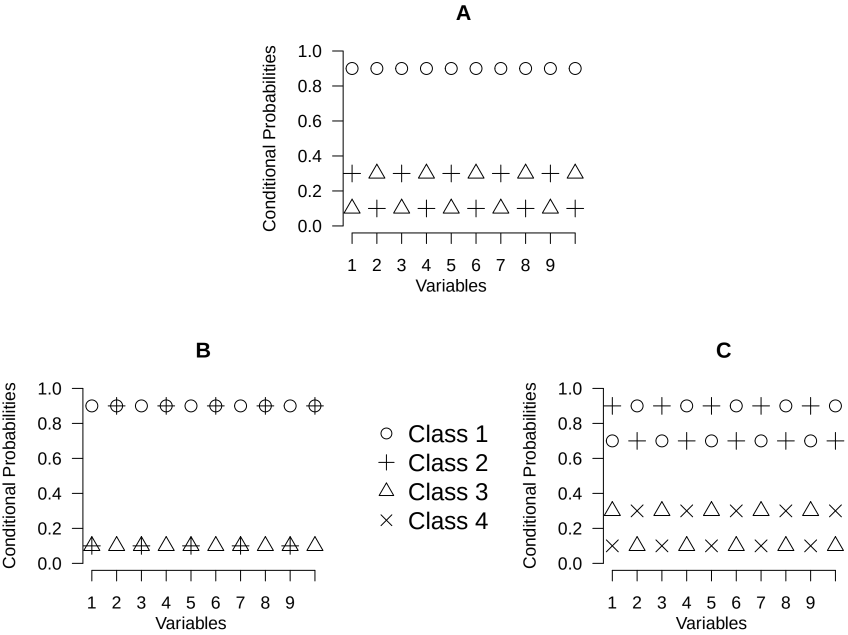

Figure 3.2: Two examples of three classes and one of four classes.

of classes nor the encountered class-specific response probabilities will correspond to what could be expected. Hence, using only binary splits is not appropriate in this case and a ternary split, or 3-class LC model, as shown in Figure 3.4, should be preferred. Note that this is not a LCT yet, but further splitting one of the three classes results in a tree structure.

The third configuration in Figure 3.2 (Panel C) contains four classes. Applying a binary split in this situation results in a child node combining classes 1 and 2 with response probabilities of 0.8 and another node combining classes 3 and 4 with response probabilities of 0.2 on the other side. Each of these combinations is split further, resulting in the tree structure of Figure 3.3 with both the expected number of classes and the appropriate conditional response probabilities.

While these illustrative examples are somewhat artificial, in real data applications with a larger number of classes, other types of class configurations may arise in which a LCT with simple binary splits may not be the right way to go. To overcome this limitation of the current LCT models, we propose a new procedure that is somewhat intermediate between a LCT analysis and

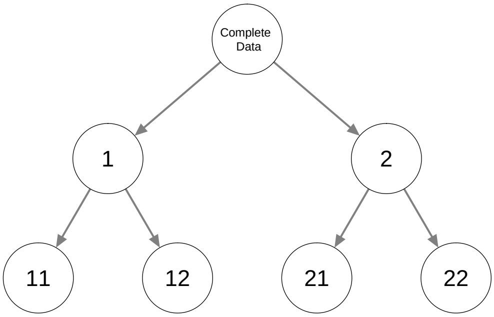

Figure 3.3: A tree structures with three binary splits.

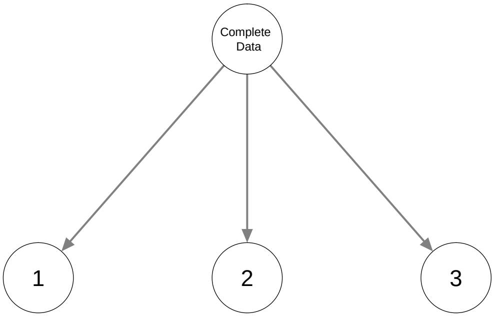

Figure 3.4: A 3-class LC model.

a standard LC analysis. Though in principle the procedure can be applied at any node of a tree, since the first split picks up the most dominant associations in the data and moreover affects most strongly the tree structure, we focus on the root node where we allow the number of child nodes to be larger than two. Various approaches can be used to decide on the number of starting classes. One option is that a researcher specifies the number of classes at the root based on theoretical grounds, and lets the binary LCT algorithm discern possible subclasses. When a priori knowledge or beliefs about the number of classes is absent, one may select the number of starting classes such that they have a clear interpretation. Note that while choosing the number of starting classes based on what is substantively meaningful ignores the statistical fit of the model, model fit is still warranted since the LCT picks up remaining associations (i.e., misfit) when classes are split up further down the tree. We also present a method for choosing the number of starting classes based on the statistical fit index. More specifically, we propose choosing the number of classes in the first split based on a relative improvement in fit measure.

The remainder of the paper is set up as follows. In the next section we discuss the basic LC model and how it can be used to build a LCT. After that we describe the measure of relative improvement in fit that we propose to determine the split size at the root, and moreover present a small simulation study on its performance in the situations depicted in Figure 3.2. Then, two empirical examples are presented illustrating how the improvement of fit measure and substantive reasoning can be used to determine the appropriate number of classes at the first split of a tree. The paper is concluded with final remarks by the authors.

3.2 Method

3.2.1 LC models

Let y_{ij} denote the response of individual i on the j^{th} categorical variable. The responses of individual i on the full set J variables is denoted by $\mathbf{y}_i$. A standard LC analysis defines a model for the probabilities of observing the various possible response patterns. Let X denote the discrete latent class variable, k denote a particular latent class, and K the number of latent classes. A LC model is specified for $P(\mathbf{y}_i)$ as follows:

$$P(\mathbf{y}_i) = \sum_{k=1}^{K} P(X = k) \prod_{j=1}^{J} P(y_{ij}|X = k). \tag{3.1}$$

Here, the probability of belonging to class k is represented by $P(X = k)$ and the probability of giving the response concerned conditional on belonging to class k is represented by $P(y_{ij}|X = k)$. The product of the class-specific response probabilities of the J variables follows from local independence assumption.

The model parameters are usually estimated by maximizing the likelihood through the EM algorithm (Dempster et al., 1977). The log-likelihood function is as follows:

$$\log L(\theta; \boldsymbol{y}) = \sum_{i=1}^{N} \log P(\mathbf{y}_i), \tag{3.2}$$

where $P(\mathbf{y}_i)$ takes the form defined in Equation (3.1), θ contains the model parameters $P(X = k)$ and $P(y_{ij}|X = k)$, and N denotes the total sample size.

3.2.2 Building a LCT

Building a LCT starts with the estimation of a standard one- and two-class model at the root node. If the two-class model is preferred, individuals are assigned to the two child classes having the root node as their parent. While the current LCT model is restricted to binary splits, below we show how to decide about a possibly larger number of starting classes. Subsequently, at the next level of the tree, the child nodes become parent nodes themselves. For each parent class, one- and two-class models are estimated, and it is decided whether a two-class model is preferred. If so, the cases belonging to the parent class concerned are assigned to the newly formed child classes, and the same procedure is repeated at the next level of the tree.

The model defined at a particular parent node is very similar to a standard LC model; i.e, it can be formulated as follows:

$$P(\mathbf{y}_i|X_{parent}) = \sum_{k=1}^{K} P(X_{child} = k|X_{parent}) \prod_{j=1}^{J} P(y_{ij}|X_{child} = k, X_{parent}), \quad (3.3)$$

where X_{parent} represents one of the parent classes at a particular level of the tree, and X_{child} represents one of the K possible newly formed child classes at the next level for the parent class concerned, with in general K equals 2. It should be noted that each child has only one parent. Hence, X_{child} actually represents $X_{child|parent}$, but for the purpose of readability, we use the shorthand X_{child} throughout this paper. Furthermore, $P(X_{child} = k|X_{parent})$ and $P(y_{ij}|X_{child} = k, X_{parent})$ represent the class proportion and the class-specific response probabilities for child class k within the parent node concerned. In other words, as in a standard LC model we define a model for $\mathbf{y}_i$, but now conditioning on belonging to a particular parent node.

As indicated above, if a split is accepted and new child classes are formed, observations are assigned to the newly formed classes based on their posterior class membership probabilities. More specifically, the posterior class membership probabilities for the K child nodes conditional on the parent node are obtained as follows:

$$P(X_{child} = k|\mathbf{y}_i; X_{parent}) = \frac{P(X_{child} = k|X_{parent}) \prod_{j=1}^{J} P(y_{ij}|X_{child} = k, X_{parent})}{P(\mathbf{y}_i|X_{parent})}. \quad (3.4)$$

However, the actual class assignment can be done in several ways, among others using modal, random, or proportional assignment rules (Dias & Vermunt, 2008). As proposed by Van der Palm et al. (2016), we use proportional class assignment in which every respondent is present at each node with a weight equal to the posterior membership probability for the node concerned.

Estimation of the LC model at the parent node X_{parent} involves maximizing the following weighted log-likelihood function:

$$\log L(\theta; \boldsymbol{y}, X_{parent}) = \sum_{i=1}^{N} w_{i,X_{parent}} P(\mathbf{y}_i | X_{parent}), \tag{3.5}$$

where $w_{i,X_{parent}}$ is the weight for person i at the parent class, which equals the posterior probability of belonging to the parent class for the individual concerned. If a split is performed, the weights for the two newly formed classes at the next level are obtained as follows:

$$w_{i,X_{child}=1} = w_{i,X_{parent}} P(X_{child} = 1 | \mathbf{y}_i; X_{parent}) \tag{3.6}$$

$$w_{i,X_{child}=2} = w_{i,X_{parent}} P(X_{child} = 2 | \mathbf{y}_i; X_{parent}). \tag{3.7}$$

In other words, a weight at a particular node equals the weight at the parent node times the posterior probability of belonging to the child node concerned conditional on belonging to the parent node. As an example, the weights $w_{i,X_1=2}$ used for investigating a possible split of class $X_1 = 2$ are constructed as follows:

$$w_{i,X_{12}} = w_{i,X=1} P(X_1 = 2 | \mathbf{y}_i, X = 1), \tag{3.8}$$

where in turn $w_{i,X=1} = P(X = 1 | \mathbf{y}_i)$. This implies:

$$w_{i,X_{12}} = P(X = 1 | \mathbf{y}_i) P(X_1 = 2 | \mathbf{y}_i, X = 1), \tag{3.9}$$

which shows that a weight at level two is in fact a product of two posterior probabilities. More details on the estimation procedure can be found in Van der Palm et al. (2016).

Construction of a LCT can be performed using standard software for LC analysis, namely by running multiple LC models with data sets containing the appropriate case weights. After each accepted split a new data set is constructed and the procedure repeats itself, which is displayed in pseudo-code in Algorithm 1. We developed an R package that automatizes these steps and which calls a LC routine – in our case version 5.1 of the Latent GOLD program (Vermunt & Magidson, 2016) – to perform the actual estimation of the LC models using the weighted data sets. This routine also provides graphical displays of the class profiles as well as of the tree structure. Thus once the tree is formed, one can investigate the discrepancies between classes at every split using profile plots. An example of a graphical representation of a LCT can be seen in Figure 3.5. To prevent the structure of the tree to be affected by the fact that classes can be permuted without changing the model fit, our R routine orders the child classes within a split based on their size in descending order.

Algorithm 1 Algorithm to construct a LCT

```
Decide on the number of classes at the first split of the tree (on the complete data)
based on the relative improvement of fit measure. Make a new data set for every
new class where each observation gets as a weight equal to its posterior probability
for the class concerned
while Splits have been made at the previous level of the tree do
    for Every new class at the previous level do
        if A split is preferred over no split then
Construct a new data set for each class and estimate 1 and 2 class models to decide
whether a further split is needed
```

3.2.3 Statistics used to define the splits.

Different types of statistics can be used to determine whether a split should be accepted or rejected. Here, we use the BIC (Schwarz, 1978), which is defined as follows:

$$BIC = -2\log L(\theta; \mathbf{y}, X_{parent}) + \log(N)P, \tag{3.10}$$

where $\log L(.)$ represents the log-likelihood at the parent node concerned, N the total sample size, and P the number of parameters of the model at hand. Thus, a split is performed if at the parent node concerned the BIC for the 2-class model is lower than the one of the 1-class model. Note that using a less strict criterion (e.g. AIC) yields the same splits as the BIC, but also possible additional splits, and thus a larger tree. In other words, depending

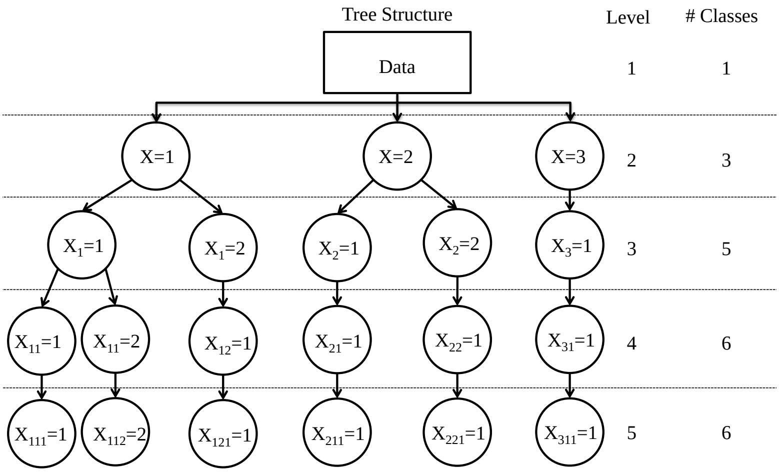

Figure 3.5: Graphical example of a LCT with
a first split into three classes.

on whether one wishes a smaller or a larger tree, a more conservative or a more liberal criterion can be used.

As explained in the introduction, in some situations, a binary split may be too much of a simplification, and one would prefer allowing for more than two classes. This is especially true for the first split of the tree, in which one picks up the most dominant features in the data. However, for this purpose, we cannot use the usual criteria like a AIC or BIC, as this would boil down to using again a standard LCT model. Instead, for the decision to use more than two classes at the first split, we propose looking at the relative improvement of fit compared to the improvement between the 1- and 2-class model. When using the log-likelihood value as the fit measure, this implies assessing the increase in log-likelihood between, say, the 2- and 3-class model and compare it to the increase between the 1- and 2-class model. More explicitly, the relative improvement between models with K and $K+1$ classes ($RI_{K,K+1}$) can be computed as:

$$RI_{K,K+1} = \frac{\log L_{K+1} - \log L_K}{\log L_2 - \log L_1}, \tag{3.11}$$

which yields a number between 0 and 1, where a small value indicates that the K-class model can be used as the first split, while a larger value indicates that the tree might improve with an additional class at the first split of the tree. Note that instead of an increase in log-likelihood, in Equation 3.11 one may use other measures of improvement of fit, such as the decrease of the

BIC or the AIC.

To get an indication of the performance of the $RI_{K,K+1}$, we run a small simulation study using the three scenarios discussed in the introduction and depicted in Figure 3.2. For each scenario we generated 100 data sets containing 10 dichotomous response variables for 1000 respondents and assuming equal class sizes. Results on the relative improvements from 2 to 3 classes and from 3 to 4 classes are shown via boxplots in Figure 3.6.

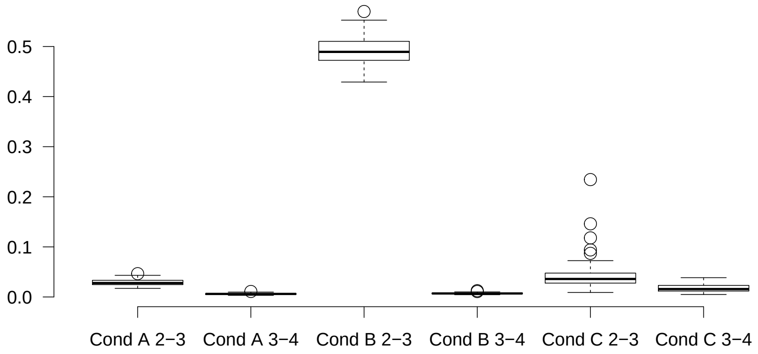

Figure 3.6: Boxplots of the improvement in fit from 2 to 3 and from 3 to 4 classes relative to the improvement from 1 to 2 classes, based on the configurations presented in Figure 3.2.

For configuration A, binary splits suffice as is shown by the always very low relative improvement when adding a third class. For configuration B, a ternary split is more suitable, which is confirmed by the high relative improvement in fit when increasing the classes from 2 to 3 obtained for every simulation replication. For configuration C, our measure indicates that a binary option suffices since the relative improvement was smaller than .10 for most of the simulation replications. Compared to the first configuration, the sampling fluctuation is somewhat larger in this configuration, which explains why a somewhat larger values were found in a small portion of the simulation replications.

3.3 Empirical examples

The proposed LCT methodology is illustrated by the analyses of two data sets which were previously studied using a standard LC model. The data set in the first example comes from a study by Owen and Videras (2008) and contains both a large number of respondents and a large number of variables, yielding a situation for which LCTs are well suited. For this data set, we compare the original LC solution by Owen and Videras (2008), the first splits of a binary LCT, and a LCT with a more appropriate number of child classes at the root using our relative improvement of fit measure. The second example concerns a very large data set in term of the number of observations from Moors and Vermunt (2007) and uses a LC model for ranking data. A LCT is very suited for this data set, as a traditional LC analysis indicates that the fit improves up to a large number of classes.

3.3.1 Social capital

Owen and Videras (2008) used the information from 14.527 respondents of several samples of the General Social Survey to construct "a typology of social capital that accounts for the different incentives that networks provide." Social capital is a construct that is plagued by "conceptual vagueness" (Durlauf & Fafchamps, 2004) and therefore Owen and Videras (2008) perform a Latent Class analysis to grasp this concept. The data set used by Owen and Videras (2008) contains sixteen dichotomous variables indicating whether respondents participate in specific types of voluntary organizations (the organizations are listed in the legend of Figure 3.7) and two variables indicating whether respondents agree with the statements "other people are fair" and

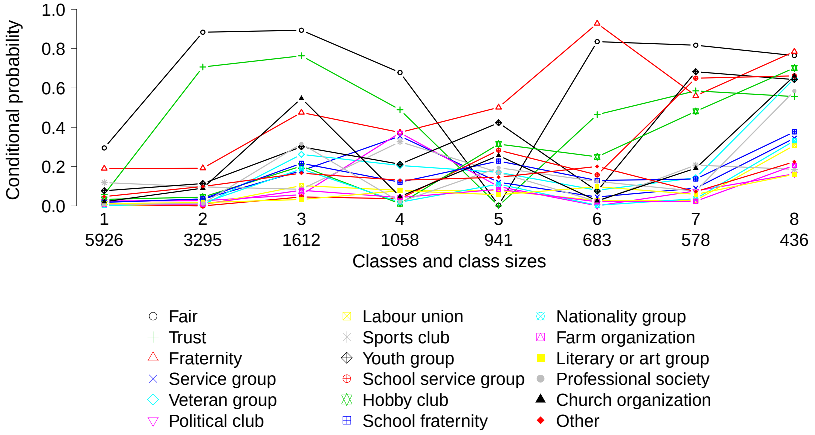

Figure 3.7: Profile plot of a standard LC analysis on social capital.

"other people can be trusted". Owen and Videras explain the inclusion of the latter two variables by stating that social capital is a multidimensional concept which embeds multiple manifestations of civic engagement as well as trust and fairness. Using the BIC, Owen and Videras selected a model with eight classes, while allowing for one local dependency, namely between the variables fraternity and school fraternity. The 8-class original solution by Owen and Videras (2008) is displayed in Figure 3.7[*], with the size of the classes displayed on the x-axis.

The classes retrieved by Owen and Videras (2008) are quite difficult to interpret. Classes 1 and 2 seem to mainly differ on the variables fair and trust, while classes 2 and 3 differ on almost all variables but fair and trust. The differences between classes 1 and 3 are subsequently a lot harder to pinpoint and this becomes increasingly difficult when including the other classes in the comparisons. Note furthermore that various of the classes contain have small class proportion (classes 4 to 8 each contain less than 10% of the observations). To facilitate the interpretation of a classification of social capital, a LCT is built with this data.

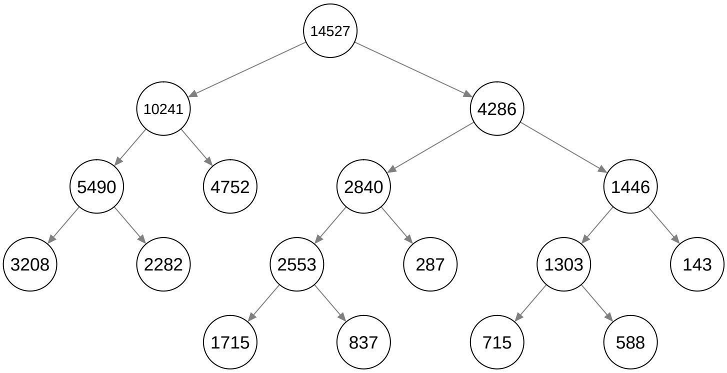

Figure 3.8: Layout of a LCT starting with a two-class split on the social capital data set.

The layout and class sizes[†] of a binary LCT based on the data of Owen and Videras (2008) is shown in Figure 3.8. The fifth and final level of the tree

[*] The exact conditional probabilities of the LC model and the LCTs on social capital can be found in Appendix A.

[†] Every split should sum up to a the class size of its parent node. However, because the allocation is carried out on the basis of the posterior probabilities, the class sizes are not integers. For convenience, these numbers have been rounded, which causes slight deviations where the sum of two child nodes does not exactly add up to the parent node.

consists of nine classes (every class which is not split further from a certain level, is taken passed as it is to a next level).

The first two levels of the binary LCT can be closer examined in their profile plots in Figure 3.9. The top panel shows the first split, which indicates that the probabilities on all variables are higher for class 2 than for class 1. So basically the first split divides the sample based on general social capital, where class 1 contains respondents with low social capital and class 2 respondents with high social capital. Within each of these groups a pessimistic (classes 11 and 22) and optimistic (classes 12 and 21) social capital group seems to be present, as these groups are split mainly on the variables fair and trust. The fact that both splits at this level are mainly due to these two variables indicates that there is a large amount of residual association between these variables within the two classes formed at the root. Hence, a tree starting with more classes at the first split may perhaps be better suited.

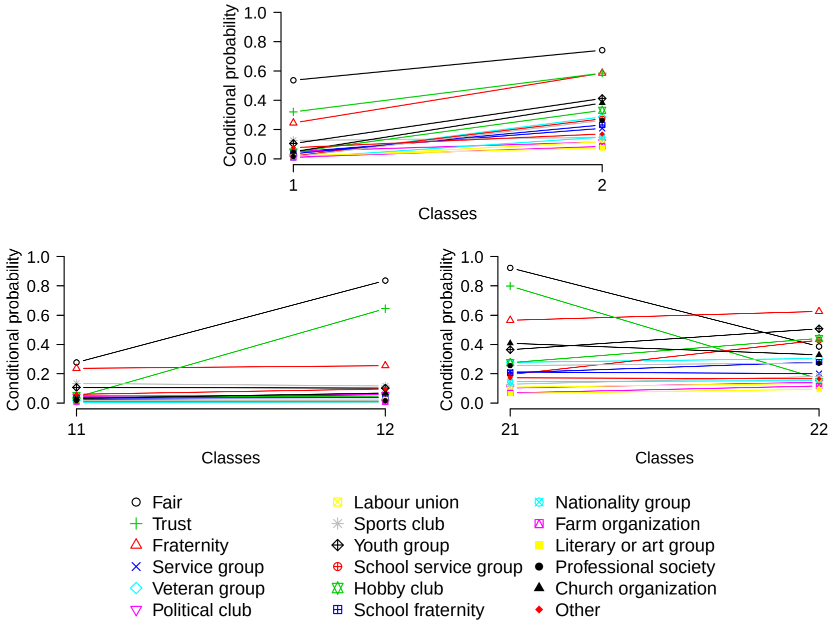

Figure 3.9: Profile plots of the first two levels of a LCT on social capital with only 2-class splits. Conditional response probabilities of the 18 items are shown on the y-axis and different (sub)classes are shown on the x-axis.

To decide on the number of classes at the root of the tree, multiple standard LC models with increasing number of classes are estimated. The fit statistics and the relative improvement of the fit statistics are shown in Table

3.1. The relative fit improvement is about 20% when expanding a model from 2 to 3 classes, compared to the improvement in fit when expanding from 1 to 2 classes. Adding more classes improves the fit marginally, indicating that a root size of three classes may be used. The complete LCT obtained by starting with three classes is shown in Figure 3.10, with the class sizes displayed for every node of the tree. For every final node it holds that, according to the BIC, a 1-class model is preferred to a 2-class model.

Table 3.1: Fit statistics and their relative improvement of the social capital data.

	$logL$	P	BIC	AIC	R_{LL}	R_{BIC}	R_{AIC}
1	-94204	18	188581	188444			
2	-89510	37	179376	179095	1.000	1.000	1.000
3	-88501	56	177539	177115	0.215	0.199	0.212
4	-88117	75	176952	176383	0.082	0.064	0.078
5	-87826	94	176553	175840	0.062	0.043	0.058
6	-87619	113	176321	175464	0.044	0.025	0.040
7	-87425	132	176114	175113	0.041	0.022	0.038
8	-87322	151	176090	174945	0.022	0.003	0.018
9	-87234	170	176098	174808	0.019	-0.001	0.015

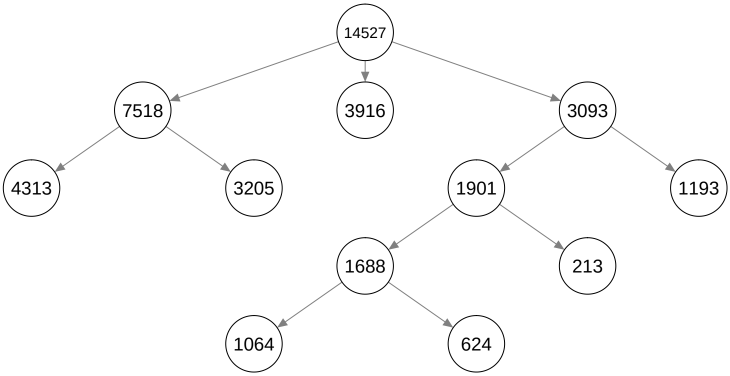

Figure 3.10: Layout of a LCT starting with a three-class split on the social capital data.

The profile plots for the splits of the LCT with three initial classes are shown in Figure 3.11, while the exact probabilities can be found in Appendix B. At first split, the first class has a low probability on all variables, the second class displays a low probability on participation in all voluntary organizations and very high probabilities on the variables fair and trust, and the third class displays relative high probabilities on participation in the voluntary organizations and rather high probabilities for fair and trust. Subsequently, the

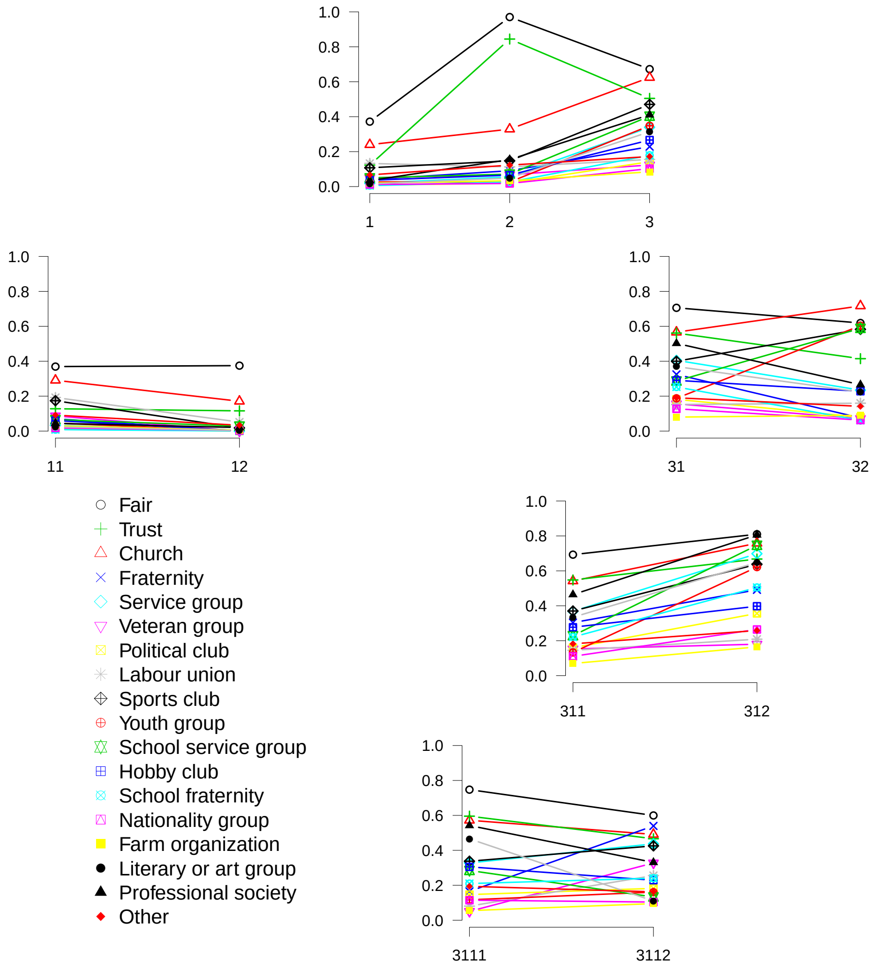

Figure 3.11: Profile plots of a LCT with a root of three classes on social capital.

first and third class are split further, while the second is not. The first class is split in a class with low and very low probabilities on all variables, while the third class is split in two classes with preferences for different voluntary organizations (e.g., a high probability for being part of a professional organization in class 31 versus a high probability for being part of a youth group in class 32). Subsequently, class 31 is split further into classes 311 and 312, which seem to differ mainly in participation in all voluntary organizations. The final split yielding classes 3111 and 3112 results in classes which differ again in preferences for different voluntary organizations (e.g, a high probability for being part of a literary or art group in class 3111 versus a high probability for being part of a fraternity in class 3112).

The original solution of eight classes by Owen and Videras (2008) can be compared with the LCT with three initial classes. Note the resemblance between the first classes of the LCT and the standard LC model. The relation between the fully binary LCT and standard LC analysis solutions is less clear, though there are also similarities. For instance, LCT-class 21 is rather similar to standard LC analysis class 2. Similarities in the results of the LCT and standard LC analysis are expected, though the goal of a LCT is not to resemble the standard LC analysis result. A great advantage of the LCT is that the classes can be interpreted stepwise, as first the classes at the first level of the tree can be interpreted and subsequently the classes at lower levels. Moreover, it offers the possibility to make a decision on the number of classes based on substantive reasons. Hence, splits at lower levels which are of no substantive interest can be ignored. For instance, the distinction between classes 11 and 12, which differ mainly in the degree of low participation in voluntary groups may be of less interest, as it reflects subtle quantitative differences rather than qualitative differences. In such a case, class 1 can be used in the final classification instead of classes 11 and 12.

3.3.2 (Post-)Materialism

The study by Moors and Vermunt (2007) used the answers of 21468 respondents participating in the 1990 European Values Survey on three questions of meant to validate the measurement of (post-)materialism as proposed by Inglehart (1971). Each item contained four aims of a country and respondents were to determine which aim should have the highest priority and which one should have the second highest priority in their opinion. The response options of the three item can be seen in Table 3.2.

Moors and Vermunt (2007) used a latent class discrete choice model for their study, as every respondent gave two ranked responses per item. A latent class discrete choice model is quite similar to a traditional latent class model as depicted in Equation (3.1). For response pattern s, with the first and second response on an item denoted as by a_1s and a_2s respectively, a discrete choice model has the form of:

$$P(y_s) = \sum_{k=1}^{K} P(X=k) \prod_{j=1}^{J} P(y_{1j} = a_{1s}, y_{2j} = a_{2s} | X = k). \tag{3.12}$$

With a LCT approach this model becomes:

$$P(y_s|X_{parent}) = \sum_{k=1}^{K} P(X_{child} = k|X_{parent}) \prod_{j=1}^{J} P(y_{1j} = a_{1s}, y_{2j} = a_{2s}|X_{child} = k, X_{parent}). \tag{3.13}$$

Within a discrete choice framework the choice probabilities are parameterized, in terms of the utilities of the alternatives. In our case, for the first item, this implies that

$$P(y_{11} = a_{1s}, y_{21} = a_{2s}|X_{child} = k, X_{parent}) = \frac{\tau_{a_1k}}{\sum_{a=1}^{4} \tau_{ak}} \frac{\tau_{a_2k}}{\sum_{a\neq a_1}^{4} \tau_{ak}}. \tag{3.14}$$

A higher value of τ_{ak} indicates a higher probability that someone belonging to class k selects alternative a. Two important differences with a standard LC model are that the utilities are assumed to be equal between the first and second choices and that it should be taken into account that the first and second choice a ranking task cannot be the same, which is why the summation for the second choice is over the non-selected alternatives ($a \neq a_1$). As is usually done, we use log transformed utilities, which are logit coefficients; that is:

$$\log \tau_{ak} = \beta_{ak} \tag{3.15}$$

For identification, effects coding is used implying that the β_{ak} sums to 0 within latent class k. The larger positive β_{ak}, the more attractive alternative a for someone belonging to the class k, while the reverse applies to negative values.

The fit statistics obtained when estimating LC discrete choice models with

Table 3.2: Indicators for the latent class discrete choice model

Item A	Item B	Item C
• Maintaining a high level of economic growth.	• Maintaining order in the nation.	• A stable economy.
• Making sure the country has strong defense forces.	• Giving people more say in important government decisions.	• Progress toward a less impersonal and more human society.
• Seeing that people have more say about how things are done at their jobs and in their communities.	• Fighting rising prices.	• Progress toward a society in which ideas count more than money.
• Trying to make our cities and countryside more beautiful.	• Protecting freedom of speech.	• The fight against crime.

1 to 10 classes, as well as the corresponding relative fit improvement are reported in Table 3.3. As can be seen, the BIC and AIC values keep decreasing till 10 classes, indicating that a large number of classes should be selected based on the measures. However, the relative improvement of fit decreases rather quickly and seems to become rather small after four classes. Thus based on this measure, a LCT model with 4 starting classes seems to be suited for this data set.

Table 3.3: Fit statistics and their relative improvement of the Discrete Choice data.

	$logL$	P	BIC	AIC	R_{LL}	R_{BIC}	R_{AIC}
1	-98236	9	196557	196489			
2	-95154	19	190490	190347	1.00	1.00	1.00
3	-94389	29	189056	188837	0.25	0.24	0.25
4	-93965	39	188304	188009	0.14	0.12	0.13
5	-93796	49	188060	187689	0.05	0.04	0.05
6	-93678	59	187920	187474	0.04	0.02	0.04
7	-93596	69	187853	187331	0.03	0.01	0.02
8	-93531	79	187818	187220	0.02	0.01	0.02
9	-93465	89	187782	187109	0.02	0.01	0.02
10	-93416	99	187779	187030	0.02	0.00	0.01

Besides the relative improvement of fit, other (substantive) considerations can be appropriate to decide on the number of classes at the first split of the tree. This is also what Moors and Vermunt (2007) did in the original study. They compared the two- to five-class models and concluded that four classes could be identified in which at least one item from each set is related to a particular latent class. Such substantive reasoning can also guide a decision on the number of classes, but with the LCT approach these classes can further be explored. Out of the four initial classes, two are split based on the BIC, and at the final level there is one more split. This yield a total number of seven classes at the final level of the tree, as is shown in Figure 3.12.

The estimated utilities are reported in Table 3.4. For the first class at the first level of the tree it can be seen that the high utilities for the first response option of every item, (to wit, the issues 'Maintaining a high level of economic growth', 'Maintaining order in the nation' and 'A stable economy'), shape the first class. These economic and 'maintaining order' issues made Moors and Vermunt (2007) interpret this class as a 'conservative' elite class, which stresses issues of macro-socio-economic order. For the second class the response options 'strong defense forces', 'fighting rising prices' and 'fight against crime' cluster together. These issues have been interpreted as 'typical' concerns of the lower class. The third class favors the more post-materialistic

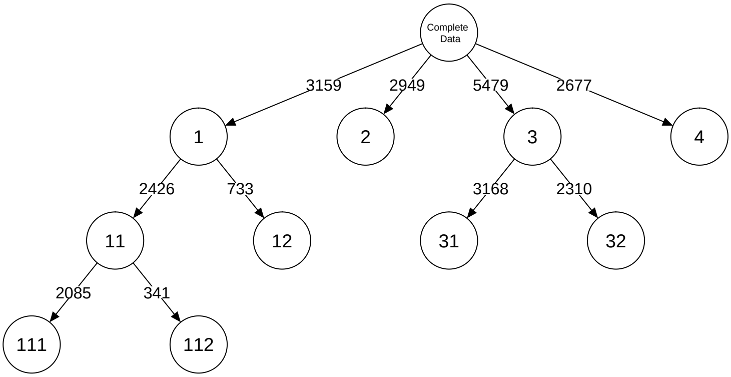

Figure 3.12: Layout and class sizes of the LCT based on the discrete choice data on (Post-)Materialism.

response options 'More say at work', 'More say in government decisions' and 'More human society'. This class is therefore also interpreted as a post-materialist class. The fourth and final class combines post-materialistic and economic issues, to wit, 'Economic growth', 'More say in government issues' and 'A stable economy'. This is interpreted as a more democratic but also macro-economic class.

These four classes at the first level are the same as those identified by Moors and Vermunt (2007) using a traditional latent class analysis. However, the tree extension allows obtaining a more detailed picture regarding the more subtle variation within these four classes. The first thing that stands out is that only classes 1 and 3 are split into subclasses. The first, so-called 'conservative' elite, class splits in two classes which differ mainly in how much

Table 3.4: Logits of the latent class discrete choice models.

Level of the tree	1				2				3	
Classes	1	2	3	4	11	12	31	32	111	112
Set A										
Economic growth	1.590	0.217	0.302	2.075	1.571	1.750	0.438	0.045	1.452	1.800
Strong defence forces	-0.992	-0.797	-2.178	-1.526	-1.525	-0.609	-2.255	-2.123	-1.311	-1.813
More say at work	0.009	0.561	1.662	0.440	0.456	-0.514	1.490	2.088	0.667	0.209
Beautiful cities	-0.606	0.019	0.213	-0.989	-0.502	-0.626	0.326	-0.010	-0.808	-0.196
Set B										
Maintaining order	1.678	0.160	-0.500	-0.652	1.796	1.581	-0.299	-0.962	1.983	1.685
More say	-0.924	-0.334	0.774	0.617	-0.839	-0.996	0.476	1.532	-0.807	-0.852
Fight rising prices	-0.521	0.470	-0.893	0.024	-0.886	-0.154	-1.183	-0.618	-0.696	-1.292
Freedom of speech	-0.233	-0.297	0.619	0.010	-0.071	-0.431	1.005	0.048	-0.480	0.460
Set C										
Stable economy	1.467	0.050	-0.591	1.638	1.356	1.619	-0.663	-0.488	1.367	1.353
Humane society	-0.484	-0.206	1.050	-0.223	-0.366	-0.648	0.918	1.314	-0.222	-0.577
Ideas count	-1.415	-0.658	0.112	-1.102	-1.365	-1.476	0.188	-0.014	-1.465	-1.234
Fight against crime	0.432	0.814	-0.570	-0.313	0.375	0.504	-0.443	-0.813	0.319	0.459

they (dis)like ’more say at work’ on the first item and how much they dislike ’strong defense forces’ on the first item and ’fighting rising prices’ on the second item. The third class at the first level, labelled the post-materialist class, is split into two classes which mainly differ in the importance attributed to ’protecting freedom of speech’ and ’giving people more say in important government decisions’. Hence, here one can distinguish two groups that differ in their preference for the post-materialistic aspects. At the final level of the tree the so-called ’conservative’ elite class that focused mainly on economic growth is split further. This split is based mainly on difference on the first and second item, where class 111 has a stronger preference for the options ’Strong defense forces’ and ’More say at work’ on item one and the option ’fighting rising prices’ on item two, and class 112 has a stronger preference for the option ’beautiful cities and countryside’ on item one and ’protecting freedom of speech’.

To summarize, the tree starts with four branches which correspond with the four classes of the original solution by Moors and Vermunt (2007), and subsequently yields five subclasses spread over two branches. The final result at the lowest level of the tree consists of 7 classes, but it is possible to decide on the most interesting number of classes of LCT with substantive reasoning. For instance, if for a particular study specific clusters of an elite class are of interest, but not a division of the post-materialistic class, classes 31 and 32 can be replaced by class 3.

3.4 Discussion

The LCT models approach discussed in this paper provide an alternative approach to LC analysis, in which a stepwise procedure is used to build a meaningful cluster model for the data set at hand. LCT models are especially useful when standard LC models would yield a large number of classes with mutual differences which are difficult to interpret. Because the restriction of the current LCT to binary splits can be problematic, we proposed a modification allowing for a larger number of child classes at the root of the LCT. We introduced a relative improvement of fit measure to decide about the number of classes, which turned out to work well in our small simulation study. We illustrated the new approach using two empirical examples, in which the relative improvement of fit measure indicated that one should use three and four starting classes, respectively. For the first example, we also compared trees starting with 2 and 3 classes, and showed that the latter yielded a much more easily interpretable clustering.

While in the current paper, the option of using non-binary splits has been applied only to the first split of the LCT, in principle it could also be used at the next levels of a tree. For instance, in the first example on social capital, both class 1 and 3 could be split into more than two classes. Based on the BIC this would be three and six classes, respectively. Rather than using the BIC, it may be possible to adapt our measure of relative improvement for this purpose, for instance, by comparing the improvement of fit with the one at the first split or with the one within the branch at hand. Because the number of classes at the splits can strongly affect the outcome of a LCT analysis, we recommend deciding this separately for every split, starting with the first split. Note that at lower levels of the tree more substantive information about the branch is already available which can be used to guide the decision regarding the number of subclasses.

The LCT models described in this paper are somewhat similar to the LC factor models proposed by Magidson and Vermunt (2001). For example, a tree with binary splits at the first and second level resembles a LC factor model with 2 dichotomous latent factors. However, in LC factor models not only the number of factors can be increased, but also the number of categories of the factors. While this is similar to increasing the number of subclasses in a split as discussed in this paper, an important difference is that the multiple classes corresponding to the same factor are restricted to be ordered. It may be worth investigating whether such an approach – in which the number of classes is increased but at the same type the classes are restricted to be ordered – is useful in the context of a LCT models as well. For instance, in our example on social capital, one may wish to force the splits at the first and second level to represent different dimensions, using possibly more than two classes. In such a case, it would make sense to apply a LC factor like approach at these splits of the LCT.

In this paper, we used the BIC to decide whether or not to stop the splitting process of the classes. While the BIC has been shown to perform well for standard LC analysis (Nylund et al., 2007), various other model selection criteria are available, such as the integrated classification likelihood (Biernacki, Celeux, & Govaert, 2000). Their strictness influences the probability to start a new branch within a LCT, implying that the choice for the decision criterion can affect the bottom of the tree significantly. Whereas we used the standard maximum likelihood method for the estimation of the submodels forming a

LCT, it may be worth considering other estimation procedures, such as the recently proposed minimum ϕ-divergence estimation method (Felipe, Miranda, & Pardo, 2015).

Summarizing, it can be stated that various options are available for deciding on the size of the splits of a LCT. In a purely exploratory analysis, the proposed relative improvement of fit measure seems to be a useful tool for deciding about the number of starting classes, while in other situations one may wish to base this decision on content information. The form of the tree and thus the composition of the classes will therefore be subject to the available information and requirements of the research question at hand. There are many ways to derive a clustering from a data set, and it is best to assume that there is no particular method which is correct in all situations (Hennig, 2015). In other words, we do not want to claim that the LCT approach will always yield the best or the true clusters, but this is often also unlikely for a standard LC analysis. In practice, a researcher may start with a standard LC analysis, and switch to our LCT approach when encountering difficulties in deciding about the number of classes or interpreting the differences between a possibly large number of classes.

Chapter 4

Building Latent Class Growth Trees

Abstract

Researchers use latent class growth (LCG) analysis to detect meaningful subpopulations that display different growth curves. However, especially when the number of classes required to obtain a good fit is large, interpretation of the encountered class-specific curves may not be straightforward. To overcome this problem, we propose an alternative way of performing LCG analysis, which we call LCG tree (LCGT) modeling. For this purpose, a recursive partitioning procedure similar to divisive hierarchical cluster analysis is used: classes are split until a certain criterion indicates that the fit does not improve. The advantage of the LCGT approach compared to the standard LCG approach is that it gives a clear insight into how the latent classes are formed and how solutions with different numbers of classes relate. The practical use of the approach is illustrated using applications on drugs use during adolescence and mood regulation during the day.

This chapter is in press at Structural Equation Modeling as Van den Bergh, M. & Vermunt, J.K. Building Latent Class Growth Trees https://doi.org/10.1080/10705511.2017.1389610

4.1 Introduction

Longitudinal data are used by social scientists to study development of behaviors or other phenomena. The analysis will often be done with latent growth curve models (MacCallum & Austin, 2000), with the aim to assess inter-individual differences in intra-individual change over time (Nesselroade, 1991). The typical growth model can be described as a multilevel model (Raudenbush & Bryk, 2002), in which the intercept and slopes of the time variables are allowed to vary across individuals. This heterogeneity is captured using random effects, which are basically continuous latent variables (Jung & Wickrama, 2008). This approach assumes that the growth trajectories of all individuals can be appropriately described by a single set of the growth parameters, and thus that all individuals come from a single population. Growth mixture modeling relaxes this assumption by allowing for differences in growth parameters across unobserved subpopulations; that is, each latent class has a separate growth model. However, fully unrestricted growth mixture models are seldom used in practice, in part due to frequent estimation problems, as well as the preference for simpler, restricted models. Probably the most widely used form of growth mixture modeling is Latent Class Growth (LCG) analysis, whereby the variances and covariances of the growth factors within classes are fixed to zero (Jones, Nagin, & Roeder, 2001; Nagin & Land, 1993). This assumes that all individuals within a class follow the same trajectory and thus that there is no residual heterogeneity within classes.

When a LCG model is applied, two key modeling decisions need to be made; that is, on the number of classes and on the shape of the class-specific trajectories. In general, the decision on the number of classes is of more importance than the decision on the shape of the trajectory of each class as long as the shape is flexible enough (Nagin, 2005). Typically, researchers estimate LCG models with different numbers of classes and select the best model using likelihood-based statistics, usually with information criteria like AIC or BIC, which weigh model fit and complexity. Although there is nothing wrong with such a procedure, in practice it is often perceived as being problematic, especially when the model is applied with a large data set; that is, when the number of time points and/or the number of subjects is large. One problem occurring in such situations is that the selected number of classes may be rather large (Francis, Elliott, & Weldon, 2016). This causes the class trajectories to pick up very specific aspects of the data, which might not be interesting for the research question at hand. Moreover, these specific trajectories are hard to interpret substantively and compare to each other. A second problem results from the fact that usually one would select a different number of classes depending on the model selection criterion used. Because of this, one may wish to inspect multiple solutions, as each of them may reveal specific relevant features in the data. However, it is often unclear how solutions with different numbers of classes are connected, making it very unclear to see what a model with more classes adds to a model with less classes.

To circumvent the issues mentioned above, it is most convenient to have models with differing numbers of classes that are substantively related; in other words, a model with $K+1$ classes is a refined version of a model with K classes, where one of the classes is split in two parts. Such an approach would result in a hierarchical structure, comparable to hierarchical cluster analysis (Everitt et al., 2011) or regression trees (Friedman et al., 2001). Van der Palm et al. (2016) developed an algorithm for hierarchical latent class analysis that can be used for this purpose. While they focused on density estimation, with some adaptations their algorithm has also been used to build so called latent class trees for substantive interpretation (Van den Bergh et al., 2017). In this paper, this procedure will be extended to the longitudinal framework to construct Latent Class Growth Trees (LCGT).

With LCGT analysis a hierarchical structure is imposed on the latent classes by estimating 1- and 2-class models on a ‘parent’ node, which initially comprised the full data. If the 2-class model is preferred according to a certain information criterion, the data is split into ‘child’ nodes and separate data sets are constructed for each of the child nodes. The split is based on the posterior class membership probabilities; hence, the data patterns in each new data set will be the same as the original data set, but with weights equal to the posterior class membership probabilities for the child class concerned. Subsequently, each new child node is treated as a parent and it is checked again whether a 2-class model provides a better fit than a 1-class model on the corresponding weighted data set. This procedure continues until no node is split up anymore. Because of this sequential algorithm, the classes at different levels of the tree can be substantively related, since child classes are subclasses of a parent class. Therefore, LCGT modelling allows for direct interpretation of the relationship between solutions with different numbers of classes, while still retaining the same statistical basis.

The remainder of the paper is set up as follows. In the next section, we discuss the basic LCG model and show how it can be used to build a LCGT. Also split criteria and guidelines for deviating from a binary split at the root of the tree will be discussed, together with an entropy measure for the post-hoc evaluation of the quality of splits. Two empirical data sets are used to illustrate LCGT analysis. The paper concludes with final remarks by the authors.

4.2 Method

4.2.1 Latent Class Growth models

Let y_{it} denote the response of individual i at time point t, T_i the number of measurements of person i, and $\mathbf{y}_i$ the full response vector of person i. Moreover, let X be the discrete latent class variable, k a particular latent class, and K the number of latent classes. A LCG model is, in fact, a regression model for the responses y_{it}, where time variables are used as predictors and where intercept and slope parameters differ across latent classes. We will define the LCG model within the framework on the generalized linear model, which allows dealing with different scale types of the response variable (Muthén, 2004; Vermunt, 2007).

Let $E(y_{it}|X = k)$ denote the expected value of the response at time point t for latent class k. After an appropriate transformation $g(\cdot)$, which mainly depends on the measurement level of the r esponse variable, $E(y_{it}|X = k)$ is modelled as a linear function of time variables. The most common approach is to use polynomial growth curves, which yields the following regression model for latent class k:

$$g[E(y_{it}|X = k)] = \beta_{0k} + \beta_{1k} \cdot t + \beta_{2k} \cdot t^2 + ... + \beta_{sk} \cdot t^s \tag{4.1}$$

The choice of the degree of the polynomial (the value of s) is usually an empirical matter, though polynomials of degree larger than three are seldom used. Recently, Francis et al. (2016) proposed an alternative approach involving the use of baseline splines in LCG models.

To complete the model formulation for the response vector $\mathbf{y}_i$, we have to define the form of the class-specific densities $f(y_{it}|X = k)$, which could be univariate normal for a continuous response, binomial for a binary response, etc.. The response density for class k is a function of the expected value $E(y_{it}|X = k)$ and for continuous variables also of the residual variance. The LCG model for $\mathbf{y}_i$ can now be defined as follows:

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} P(X = k) \prod_{t=1}^{T_i} f(y_{it}|X = k), \tag{4.2}$$

where the size of class k is represented by $P(X = k)$. A graphical representation of a LCG model with $K = 3$ can be seen in Figure 4.1.

The model estimates (the β parameters and class sizes) can be obtained by maximizing the following log-likelihood function:

$$\log L(\theta; \mathbf{y}) = \sum_{i=1}^{N} \log f(\mathbf{y}_i), \tag{4.3}$$

where $f(\mathbf{y}_i)$ takes the form defined in Equation (4.2) and N denotes the total sample size. Maximization is usually achieved through an EM algorithm (Dempster et al., 1977), possibly combined with a Newton-type algorithm (Vermunt & Magidson, 2013).

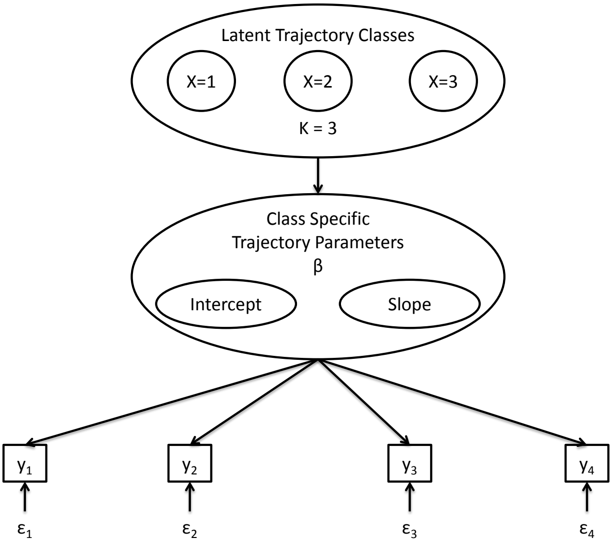

Figure 4.1: Graphical representation of a LCG model with three trajectory classes.

After selecting a particular model, individuals may be assigned to latent classes based on their the posterior class membership probabilities. Using the Bayes theorem, these probabilities are obtained as follows:

$$P(X = k|\mathbf{y}_i) = \frac{P(X = k)\prod_{t=1}^{T_i} f(y_{it}|X = k)}{f(\mathbf{y}_i)}. \tag{4.4}$$

4.2.2 Latent Class Growth Tree models

Using an algorithm similar to the algorithm developed by Van der Palm et al. (2016) for divisive latent class analysis, a LCG model can also be constructed in a tree form. Such a LCGT has the advantages that increasing K classes to $K+1$ classes results in directly related classes. This is because newly formed classes are obtained by splitting one of the K classes. Due to this direct relation, models with different numbers of classes can be substantively related, while still retaining the same statistical basis. Below we first describe the algorithm for constructing a LCGT in more detail, and subsequently discuss various statistics that can be used during this process.

A LCGT consists of parent and child nodes. Every set of child nodes is based on one parent node and the first parent node consists of the root node containing the complete data set. At each parent node, standard LCG models are used and its child nodes are the classes assessed with the selected parent model. At the next level of the tree, these child nodes, in their turn, become parent nodes, and conditional on each new parent node a new set of LCG models is defined. This process continues until a stopping criterion is reached, for example, when the BIC does no longer decrease when splitting.

The basic equations of the growth curves of a LCGT model do not differ from those of a standard LCG model (e.g., Equation 4.1). The fact that the LCGT model is based on LCG models at parent nodes can be formulated as follows:

$$Pf(\mathbf{y}_i|X_{parent}) = \sum_{k=1}^{K} P(X_{child} = k|X_{parent}) \prod_{t=1}^{T} f(y_{it}|X_{child} = k, X_{parent}), \tag{4.5}$$

where X_{parent} represents the parent class at level l and X_{child} represents one of the K possible newly formed classes at level $l+1$, with in general K being 2. Furthermore, $P(X_{child} = k|X_{parent})$ represents the size of a class, given the parent node, while $f(y_{it}|X_{child} = k, X_{parent})$ represents the class-specific response density at timepoint t, given the parent class. In other words, as in a standard LCG analysis, a model for $\mathbf{y}_i$ is defined, but now conditioned on belonging to the parent class concerned.

Estimation of the LCG model at the parent node X_{parent} involves maximizing the following weighted log-likelihood function:

$$\log L(\theta; \mathbf{y}, X_{parent}) = \sum_{i=1}^{N} w_{i,X_{parent}} P(\mathbf{y}_i | X_{parent}), \tag{4.6}$$

where $w_{i,X_{parent}}$ is the weight for person i at the parent class, which equals this person's posterior probability of belonging to the parent class concerned. So, building a LCGT involves estimating a series of LCG model using weighted data sets.

To see how the weights $w_{i,X_{parent}}$ are constructed, let us first look at the posterior class membership probabilities for the child nodes, conditional on the corresponding parent node. Assuming a split is accepted, the posteriors are obtained as follows:

$$P(X_{child} = k | \mathbf{y}_i; X_{parent}) = \frac{P(X_{child} = k | X_{parent}) \prod_{t=1}^{T_i} f(y_{it} | X_{child} = k, X_{parent})}{P(\mathbf{y}_i | X_{parent})}. \tag{4.7}$$

As proposed by Van der Palm et al. (2016), we use a proportional split based on these posterior class membership probabilities for the K child nodes conditional on the parent node, denoted by $k = 1, 2, ..., K$. If a split in two classes is performed, the weights for the two newly formed classes at the next level are obtained as follows:

$$w_{i,X_{child}=1} = w_{i,X_{parent}} P(X_{child} = 1 | \mathbf{y}_i; X_{parent}) \tag{4.8}$$

$$w_{i,X_{child}=2} = w_{i,X_{parent}} P(X_{child} = 2 | \mathbf{y}_i; X_{parent}). \tag{4.9}$$

In other words, a weight for individual i at a particular node equals the weight at the parent node times the posterior probability of belonging to the child node concerned conditional on belonging to the parent node. As an example, the weights $w_{i,X_1=2}$ used for investigating a possible split of class $X_1 = 2$ are constructed as follows:

$$w_{i,X_{12}} = w_{i,X=1} P(X_1 = 2 | \mathbf{y}_i, X = 1), \tag{4.10}$$

where in turn $w_{i,X=1} = P(X = 1 | \mathbf{y}_i)$. This implies:

$$w_{i,X_{12}} = P(X = 1 | \mathbf{y}_i) P(X_1 = 2 | \mathbf{y}_i, X = 1), \tag{4.11}$$

which shows that a weight at level two is in fact a product of two posterior class membership probabilities.

Construction of a LCGT can be performed using standard software for LC analysis, namely by running a series of LC models with the appropriate weights. After each accepted split a new data set is constructed and the procedure repeats itself. We developed an R routine in which this process is fully automated. It calls the Latent GOLD program (Vermunt & Magidson, 2013) in batch mode to estimate 1- and 2-class models, evaluates whether a split should be made, and keeps track of the weights when a split is accepted. In addition, it creates various graphical displays which facilitates the interpretation of the LCGT (see among others Figure 4.2). A novel graphical display is a tree depicting the class-specific growth curves for the newly formed child classes (for an example, see Figure 4.4). In the trees, the name of a child class equals the name of the parent class plus an additional digit, a 1 or a 2. To prevent that the structure of the tree will be affected by label switching resulting from the fact that the order of the newly formed classes depends on the random starting values, when building the LCGT we locate the larger class at the left branch with number 1 and the smaller class at the right branch with number 2.

4.2.3 Statistics for building and evaluating the LCGT

Different types of statistics can be used to determine whether a split should be accepted or rejected. Here, we will use the BIC (Schwarz, 1978), which is defined as follows:

$$BIC = -2\log L(\theta; \mathbf{y}, X_{parent}) + \log(N)P, \tag{4.12}$$

where $\log L(.)$ represents the log-likelihood at the parent node concerned, N the total sample size, and P the number of parameters of the model at hand. Thus, a split is performed if at a parent node concerned the BIC for the 2-class model is lower than the one of the 1-class model. Note that using a less strict criterion (e.g. AIC) will yield the same splits as the BIC, but possible also additional splits, and thus a larger tree.

Special attention needs to be dedicated to the first split at the root node of the tree, in which one picks up the most dominant features in the data. In many situations, a binary split at the root may be too much of a simplification, and one would prefer allowing for more than two classes in the first split. For this purpose, we cannot use the usual criteria like a AIC or BIC, as this would boil down to using again a standard LCG model. Instead, for the decision to use more than two classes at the root node, we propose looking at the relative improvement of fit compared to the improvement between the 1- and 2-class model. When using the log-likelihood value as the fit measure, this implies assessing the increase in log-likelihood between, say, the 2- and 3-class model and compare it to the increase between the 1- and 2-class model. More explicitly, the relative improvement between models with K and $K+1$ classes ($RI_{K,K+1}$) can be computed as:

$$RI_{K,K+1} = \frac{\log L_{K+1} - \log L_K}{\log L_2 - \log L_1}, \tag{4.13}$$

which yields a number between 0 and 1, where a small value indicates that the K-class model can be used as the first split, while a larger value indicates that the tree might improve with an additional class at the root of the tree. Note that instead of an increase in log-likelihood, in Equation 4.13 one may use other measures of improvement of fit, such as the decrease of the BIC or the AIC.

The BIC and $RI_{K,K+1}$ statistics are used to determine whether and how splits should be performed. However, often we are also interested in evaluating the quality of splits in terms of the amount of separation between the newly formed classes; that is, to determine how different the classes are. In other words, is a split substantively important or not. This is also relevant if one would like to assign individuals to the classes resulting from a LCGT. Note that the assignment of individuals to the two child classes is more certain when the larger of the posterior probabilities $P(X_{child} = k|\mathbf{y}_i; X_{parent})$ is closer to 1. A measure to express this is the entropy; that is,

$$Entropy(X_{child}|\mathbf{y}) = \sum_{i=1}^{N} w_{i|X_{parent}} \sum_{k=1}^{2} -P(X_{child} = k|\mathbf{y}_i; X_{parent}) \log P(X_{child} = k|\mathbf{y}_i; X_{parent}). \tag{4.14}$$

Typically $Entropy(X_{child}|\mathbf{y})$ is rescaled to lie between 0 and 1 by expressing it in terms of the reduction compared to $Entropy(X_{child})$, which is the entropy computed using the unconditional class membership probabilities $P(X_{child} = k|X_{parent})$. This so-called $R^2_{Entropy}$ is obtained as follows:

$$R^2_{Entropy} = \frac{Entropy(X_{child}) - Entropy(X_{child}|\mathbf{y})}{Entropy(X_{child})} \tag{4.15}$$

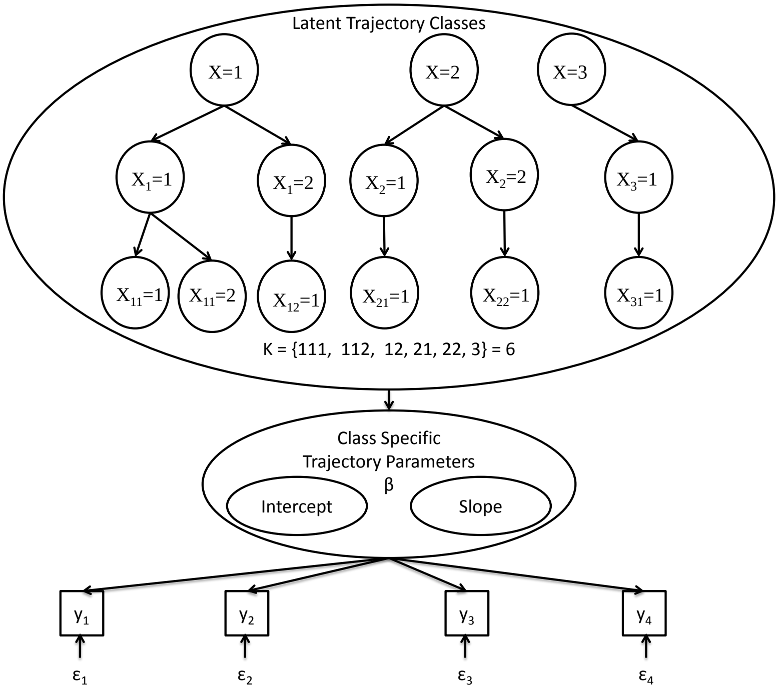

Figure 4.2: Graphical example of a LCGT model with a root of three classes.

The closer $R^2_{Entropy}$ is to one, the better the separation between the child classes in the split concerned.

4.3 Empirical examples

The proposed LCGT methodology will be illustrated by the analyses of two longitudinal data sets. The data set in the first example contains a yearly dichotomous response on drugs use collected using a panel design. The second data set contains an ordinal mood measure, recorded using an experience sampling design with eight measures per day during one week. The two data sets illustrate LCGT analyses, differing in the number of classes at their root node. For both examples, the quality of the splits will also be evaluated using the entropy-based R-squared.

4.3.1 Example 1: Drugs Use

The first data set stems from the National Youth Survey (Elliot, Huizinga & Menard, 1989). It contains nine waves, from 1976 to 1980 yearly and from 1980 to 1992 with three year intervals. The age at the first wave of the 1725 respondents (53% men and 47% women) varied between 11 and 17 years. We use age at the panel wave concerned as the time variable, which takes on values ranging from age 11 to 33. Each respondent has been observed at most nine times (on average 7.93 times). The dichotomous dependent variable of interest in our example will be whether the respondent used drugs or not during the past year.

Because the trees based on second and third degree polynomial growth curves were almost identical, the simpler one using a second degree polynomial was retained. The tree structure and the class sizes at the splits* are presented in Figure 4.3. As can be seen, there are four binary splits, which result in a total of five latent classes at the end nodes.

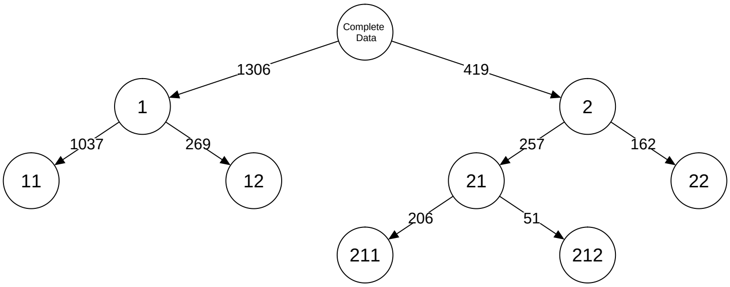

Figure 4.3: Layout of a LCGT with
a root of two classes on drugs use over age.

To determine whether it would be better to increase the number of classes at the root of the tree, we can look at the relative improvement in fit of models with more than 2 classes according to the likelihood, BIC, and AIC as reported in Table 4.1. As can be seen, the relative improvement with a third class is around 10%. As this is quite low, we retain the tree with a binary split at the root.

To interpret the encountered classes, the growth curves can be plotted for the two newly formed classes at each node of the tree. This is displayed in Figure 4.4. As can be seen, the first split results in a class with a low

*Every split should sum up to the class size of its parent node. However, because the allocation is carried out on the basis of the posterior probabilities, the class sizes are not integers. For convenience, these numbers have been rounded, which causes slight deviations where the sum of two child nodes does not exactly add up to the parent node.

Table 4.1: Likelihood, number of parameters, BIC, AIC, and relative improvement of the fit statistics of a traditional LC growth model with 1 to 6 classes.

	$\log L$	P	BIC	AIC	$RI_{\log L}$	RI_{BIC}	RI_{AIC}
1	-5089	3	10200	10183			
2	-4246	7	8543	8505			
3	-4156	11	8394	8334	0.106	0.090	0.102
4	-4086	15	8284	8202	0.083	0.067	0.079
5	-4046	19	8233	8129	0.048	0.031	0.043
6	-4028	23	8228	8102	0.021	0.003	0.016

probability to use drugs (class 1) and a class with a high probability to use drugs (class 2). Subsequently both of these classes are split further. Class 1 is split into class 11 with a very low probability of using drugs (on average 0.01%) and class 12 with a low probability during the first few years, but with a slight increase from age 20 to 33. Class 2 is split into class 21 and 22, which mainly differ in the moment at which the probability of drugs use is the highest: Respondents of class 21 start using drugs a few years earlier than respondents of class 22. Finally, class 21 is split further, where class 211 has a moderate probability (around 0.6) to use drugs at an early age, but this probability also quickly declines. Class 212 has a very high probability (around 0.95) to start using drugs at an early age and this probability stays quite constant up to age 25.

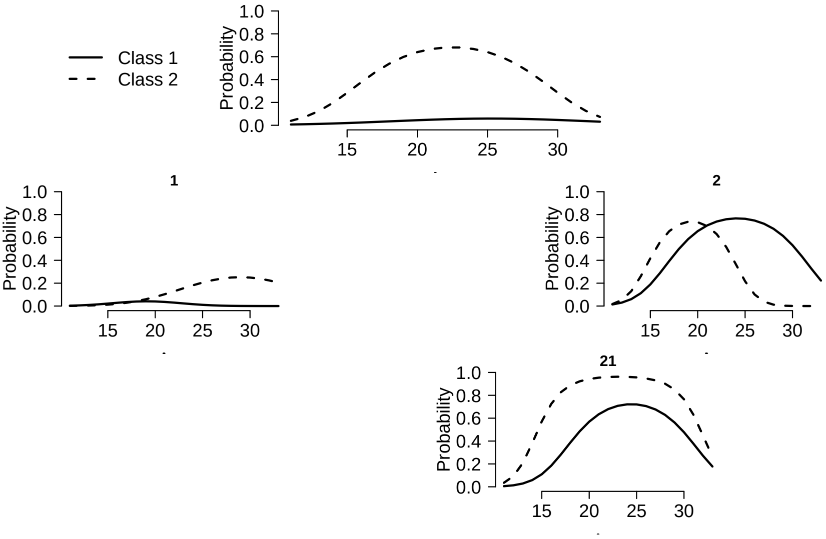

Figure 4.4: LCGT with a root of 2 classes on drug use over age.

The $R^2_{Entropy}$ values confirm what could also be seen from the depicted growth curves: The first split on the complete data set shows a large difference between the two classes with a $R^2_{Entropy}$ of 0.746. Furthermore, classes 11 and 12 are quite similar with a $R^2_{Entropy}$ of 0.268, whereas the differences between classes 21 and 22 and between classes 211 and 212 are substantial (the $R^2_{Entropy}$ values are 0.545 and 0.619 respectively). Hence, after the first split, the branch of class 2 contains more important additional differences than the one of class 1.

4.3.2 Example 2: Mood Regulation

The second data set stems from a momentary assessment study by Crayen, Eid, Lischetzke, Courvoisier, and Vermunt (2012). It contains 8 mood assessments per day during a period of one week among 164 respondents (88 women and 76 men, with a mean age of 23.7, SD = 3.31). Respondents answered a small number of questions on a handheld device at pseudo-random signals during their waking hours. The delay between adjacent signals could vary between 60 and 180 minutes (M [SD] = 100.24[20.36] minutes, min = 62 minutes, max = 173 minutes). Responses had to be made within a 30-minute time window after the signal, and were otherwise counted as missing. On average, the 164 participants responded to 51 (of 56) signals (M [SD] = 51.07 [6.05] signals, min = 19 signals, max = 56 signals). In total, there were 8374 non-missing measurements.

At each measurement occasion, participants rated their momentary mood on an adapted short version of the Multidimensional Mood Questionnaire (MMQ). Instead of the original monopolar mood items, a shorter bipolar version was used to fit the need for brief scales. Four items assessed pleasant-unpleasant mood (happy-unhappy, content-discontent, good-bad, and well-unwell). Participants rated how they momentarily feel on a 4-point bipolar intensity scales (e.g., very unhappy, rather unhappy, rather happy, very happy). For the current analysis, we focus on the item well-unwell. Preliminary analysis of the response category frequencies showed that the lowest category (i.e., very unwell) was only chosen in approximately 1% of all occasions. Therefore the two lower categories were collapsed together into one unwell category. The following analysis is based on the recoded item with three categories (conform Crayen et al. (2012)).

For the analysis, we used a LCG model based on an ordinal logit model. The time variable was the time during the day, meaning that we model the

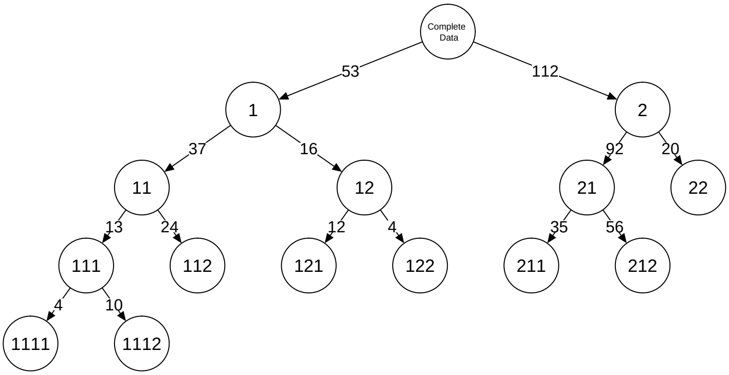

Figure 4.5: Layout of a LCGT with a root of two classes on mood regulation during the day.

mood change during the day. There was a substantial difference between a tree based on a second or a third degree polynomial, which indicates that developments are better described by cubic growth curves (see also the trajectory plots in Figure 4.7). Because there was no substantial difference between a tree based on a third or a fourth degree polynomial, a third degree polynomial was used. The LCGT model obtained with a root of two classes is quite large, with in total seven binary splits, resulting in a total of eight latent classes (Figure 4.5). A large tree already indicates that a larger number of classes at the root of the tree might be appropriate. Moreover, based on the relative improvement of the log-likelihood, BIC, and AIC (Table 4.2), it seems sensible to increase the number of classes at the root of the three.

The layout and size of the LCGT with 3 root classes can be seen in Figure 4.6 and its growth curve plots in Figure 4.7. The growth plots show that at the root of the tree, the three different classes all improve their mood during the day. They differ in their overall mood level, with class 3 having the lowest

Table 4.2: Likelihood, number of parameters, BIC, AIC, and relative improvement of the fit statistics of a traditional LC growth model with 1 to 6 classes.

	$\log L$	P	BIC	AIC	$RI_{\log L}$	RI_{BIC}	RI_{AIC}
1	-7199	4	14424	14408			
2	-6741	9	13538	13504			
3	-6578	14	13244	13191	0.355	0.333	0.347
4	-6516	19	13149	13077	0.137	0.107	0.126
5	-6471	24	13091	13001	0.097	0.065	0.085
6	-6443	29	13064	12956	0.062	0.030	0.050

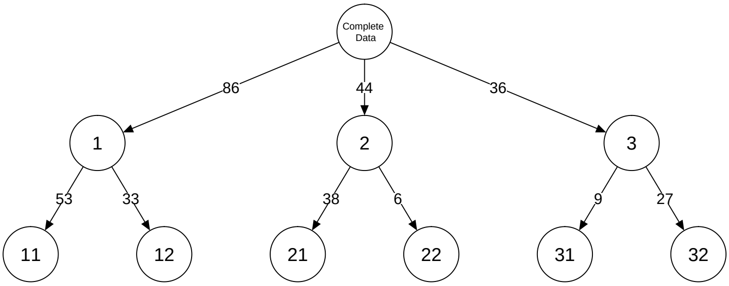

Figure 4.6: Layout of a LCGT with a root of three classes
on mood regulation during the day.

and class 2 the highest overall score. Moreover, class 1 seems to be more consistently increasing than the other two classes.

These three classes can be split further. Class 1 splits into two classes with both an average score around one, class 11 just above and class 12 just below. Moreover, the increase in class 11 is larger than in class 12. The split of class 2 results in class 21 consisting of respondents with a very good mood in the morning, a quick decrease until mid-day, and a subsequent increase. In general the mean score of class 21 is high relative to the other classes. Class 22 starts with an average mean score and subsequently only increases. The

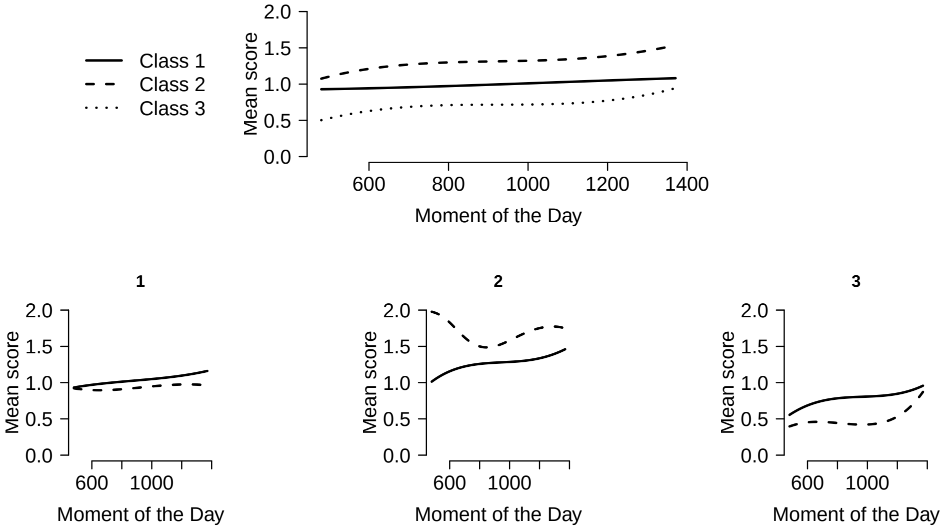

Figure 4.7: LCGT with a root of three classes
on mood regulation during the day.

splitting of class 3 results in two classes with a below average mood. Both classes increase, class 31 mainly in the beginning and class 32 mainly at the end of the day.

The $R^2_{Entropy}$ of the different splits is quite high. The root of the tree has a $R^2_{Entropy}$ of 0.889, while the $R^2_{Entropy}$ of the subsequent splits are 0.734, 0.932 and 0.897 respectively. This indicates that the differences between the subclasses 21 and 22 are larger than those between subclasses 31 and 32, while classes 11 and 12 differ the least.

4.4 Discussion

LCG models are used by researchers who wish to identify (unobserved) subpopulations with different growth trajectories using longitudinal data. However, often the number of latent classes encountered is rather large, making interpretation of the results difficult. Moreover, because solutions with different number of classes are unrelated, a substantive comparison of models with different numbers of classes is not possible, which is especially problematic when different model selection criteria point at a different optimal number of classes. To resolve these issues, we proposed using LCGT models in which the identification of the latent classes is done in a sequential manner. The constructed hierarchical tree will show the most important distinctions in growth trajectories in the first splits, and more detailed distinctions in latter splits. While we primarily used binary splits, we also showed how to decide about larger splits using relative improvement of fit measures. The latter is mainly of interest at the root of the tree. The proposed LCGT algorithm and graphical displays which are available as R code were illustrated with two empirical examples. The two illustrative examples showed that easily interpretable solutions are obtained using our new procedure.

Various extensions and variants of the proposed procedure are possible and worth to study in more detail. Whereas in the current paper we restricted ourselves to LCGTs with only binary splits after the split at the root of the tree, also at the second and next levels it may be of interest to use larger split sizes, which may result in a tree with different split sizes within branches. Because the size of the splits may strongly affect the structure of the constructed LCGT, we recommend deciding this separately per split rather than using a fully automated procedure. Note that at this stage more substantive information about the branch is available to guide a decision.

The BIC was used to decide whether or not to split a class, as it has been shown to perform well for standard LC and LCG analysis (Nylund et al., 2007). However, other measures could be used as well, where their strictness will influence the likelihood to start a new branch within a tree. Therefore, the decision criterion used can affect the bottom part of the tree significantly. Note that the lower parts are also affected by the decision to increase the number of classes at the root of the tree. Moreover, the exact choice of a criterion depends on the required specificity of the encountered growth trajectories, where a less strict criterion may be used if one wishes to see more specific classes at the bottom of the tree.

While LCG models are becoming very popular among applied researchers, the use of these models is not easy at all (Van de Schoot, Sijbrandij, Winter, Depaoli, & Vermunt, 2017). We hope that the proposed LCGT methodology will simplify the detection and interpretation of underlying growth trajectories. This does, of course, not mean that the standard LCG model is not useful anymore. In practice, a researcher may start with a standard LCG analysis, and switch to our LCGT approach when encountering difficulties in deciding about the number of classes or interpreting the differences between a possibly large number of classes.

Chapter 5

Latent Class Trees with the three-step approach

Abstract

Latent class analysis is widely used in the social sciences to cluster respondents based on a set of categorical variables. Recently, Van den Bergh et al. (2017) proposed a new type of latent class analysis, called Latent Class Tree (LCT) analysis, which provides more options to decide on the number of classes and facilitates interpretation of the classes. However, assessing the latent classes is often the first part of an analysis, as quite often the goal of a research is to relate the classes to some external variables. For this purpose the bias-adjusted three-step method is one of the preferred approaches. In this chapter we developed a bias-adjusted three-step procedure for LCT modeling, which is illustrated using applications relating social capital to demographic variables and mood regulation during the day to personality traits.

This chapter is to be submitted as Van den Bergh, M. & Vermunt, J.K. Latent Class Trees with the three-step approach

5.1 Introduction

Social scientist often use Latent Class (LC) models to cluster respondents based on their response patterns of categorical variables (Clogg, 1995; Goodman, 1974; Hagenaars, 1990; Lazarsfeld & Henry, 1968; McCutcheon, 1987). The classes represent homogeneous sub-groups of the respondents, which are interpreted based on the conditional response probabilities within a class (Muthén, 2004). Typically, researchers estimate LC models with different numbers of classes and select the best model using likelihood-based statistics which weigh model fit and complexity (e.g., AIC or BIC). Although there is theoretically nothing wrong with such a procedure, in practice it is often perceived as being problematic. Problems with LC models occur especially when the model is applied to a large data set; that is, when the number of variables and/or the number of subjects is large. One problem occurring in such situations is that the selected number of classes may be rather large. This causes the classes to pick up very specific aspects of the data, which might not be interesting for the research question at hand. Moreover, these specific classes are hard to interpret substantively and compare to each other. A second problem results from the fact that usually one would select a different number of classes depending on the model selection criterion used. Because of this, one may wish to inspect multiple solutions, as each of them may reveal specific relevant features in the data. However, it is often unclear how solutions with different numbers of classes are connected, making it very hard to see what a model with $K + 1$ classes adds to a model with K classes.

To circumvent the issues mentioned above, Van den Bergh et al. (2017) proposed the Latent Class Tree (LCT) modeling approach, which is based on an algorithm for density estimation by Van der Palm et al. (2016). LCT modeling involves imposing a hierarchical tree structure on the latent classes by estimating 1- and 2-class models on a 'parent' node, which initially comprises the full data. If the 2-class model is preferred according to a certain information criterion, the data is split into 'child' nodes and separate data sets are constructed for each of the child nodes. Subsequently, each new child node is treated as a parent and it is checked again whether a 2-class model provides a better fit than a 1-class model on the corresponding weighted data set. This procedure continues until no node is split up anymore. Because of this sequential algorithm, the classes at different levels of the tree can be substantively related, since child classes are subclasses of a parent class. Therefore, LCT modeling yields a hierarchical structure which allows for direct interpretation of the relationship between solutions with different numbers of classes, while still retaining a solid statistical basis.

However, the identification of classes is usually only the first step in a LC analysis, as researchers are often also interested in how the classes are related to one or more external variables. When doing standard LC analysis, the relation between LC membership and external variables of interest can be assessed with two different procedures; the one-step procedure in which the external variables are included in the model (Dayton & Macready, 1988; Hagenaars, 1990; Van der Heijden, Dessens, & Bockenholt, 1996; Yamaguchi, 2000) or the three-step procedure (Bakk et al., 2016; Bakk & Vermunt, 2016; Bolck et al., 2004; Vermunt, 2010). In general, the three-step approach is most often used, for several reasons. Firstly, researchers prefer separating the construction of the measurement part (in which the number of classes and their relation with the indicator variables is determined) and the development of a structural part (in which the latent classes are related to the external variables of interest). Secondly, the one-step approach also uses the external variables for the formation of latent classes, while the goal is to relate the external variables with the latent classes that would have been formed without the external variables included. This thus seems to create an unwanted circularity. Finally, the three-step approach is, especially with continuous distal outcomes, much less affected by assumptions on the class-specific conditional distribution of the external variables (Bakk & Vermunt, 2016). Hence, researchers commonly use the three-step approach. The main disadvantage of the three-step procedure as applied till about 10 years ago was that it underestimated the relation between the external variables and the latent class membership. Fortunately, recently methods have been developed to adjust for this bias (Bakk, Tekle, & Vermunt, 2013; Bolck et al., 2004; Vermunt, 2010).

For LCT modeling the three-step approach is also more logical to use than the one-step approach. This allows in the first step to build a tree whose structure and classes do not depend on the external variables. Subsequently in step 2 the individuals are assigned to the latent classes, and in step 3 the class assignments are used to investigate the relation between the latent classes and the external variable at hand. In this paper we propose a novel approach that allows investigating the relationship with covariates and distal outcomes at each split of the constructed LCT.

The content of this paper is outlined as follows. First we introduce the three steps of the proposed bias-adjusted three-step LCT modeling approach. The method is subsequently illustrated with two empirical examples. The first example is based on cross-sectional data, for which a standard LCT is build and the bias-adjusted three-step method for LCTs is used to relate some distal outcome variables to the classes. The second example is based on longitudinal data, for which a Latent Class Growth Tree (LCGT) is constructed and the bias-adjusted three-step method is used to relate covariates to the growth classes. The paper concludes with final remarks by the authors.

5.2 Method

Bias adjusted three-step LC modeling has been described among others by Vermunt (2010) and Bakk and Vermunt (2016). What we will do here is show how the three steps - building a LC model, classification and quantifying the classification errors, and bias-adjusted step-three analysis with external variables - look like in the case of a LCT model. As is shown below in more detail below, the main modification compared to a standard three-step LC analysis is that these three steps are now performed conditional on the parent class. In fact, a separate three-step analysis is performed at each node of the LCT where a split occurs.

5.2.1 Step 1: Building a LCT

The first step of bias-adjusted three-step LCT modeling involves building a LCT without inclusion of the external variables. Let $\mathbf{y}_i$ denote the response of individual i on all J variables, X the discrete latent class variable, and k a particular latent class. Moreover, subscripts p and c are used to refer to quantities of parent and child nodes, respectively. Then the 2-class LC model defined at a particular parent node can be formulated as follows:

$$P(\mathbf{y}_i|X_p) = \sum_{k=1}^{2} P(X_c = k|X_p) \prod_{j=1}^{J} P(y_{ij}|X_c = k, X_p), \tag{5.1}$$

where X_p represents the parent class at level t and X_c one of the two possible newly formed child classes at level $t+1$. In other words, as in a standard LC-model we define a model for $\mathbf{y}_i$, but now conditioning on belonging to the parent class concerned. If the 2-class model is preferred according to a certain information criterion, the data is split into 'child' nodes. This split is based on the posterior membership probabilities, which can be assessed by applying Bayes theorem to the estimates obtained from Equation (5.1):

$$P(X_c = k|\mathbf{y}_i; X_p) = \frac{P(X_c = k|X_p)\prod_{j=1}^{J} P(y_{ij}|X_c = k, X_p)}{P(\mathbf{y}_i|X_p)}. \tag{5.2}$$

For each child class a separate data set is constructed, which contain the same data as the original data set, but also the posterior membership probabilities as weights. Hereafter, each of these data sets become a parent classes themselves, and the 1-class model and the 2-class model defined in Equation (5.1) are estimated again for each newly created data set with the corresponding weights for each of the parent classes (w_p). The splitting procedure is repeated until no 2-class models are preferred anymore over 1-class models. This results in a hierarchical tree structure of classes. Within a LCT, the name of a child class equals the name of the parent class plus an additional digit, a 1 or a 2. For convenience, the classes are sorted by size, with the first class as largest class. For a more detailed description on how to build a LCT, see Van den Bergh et al. (2017).

Special attention needs to be dedicated to the first split at the root node of a LCT (or LCGT), in which one picks up the most dominant features in the data. In many situations, a binary split at the root may be too much of a simplification, and one would prefer allowing for more than two classes in the first split. For this purpose, we cannot use the usual criteria like a AIC or BIC, as this would boil down to using a standard LC model. Instead, for the decision to use more than two classes at the root node, one can look at the relative improvement in fit compared to the improvement between the 1- and 2-class model. When using the log-likelihood value as the fit measure, this implies assessing the increase in log-likelihood between, say, the 2- and 3-class model and compare it to the increase between the 1- and 2-class model. More explicitly, the relative improvement between models with K and $K+1$ classes ($RI_{K,K+1}$) can be computed as:

$$RI_{K,K+1} = \frac{\log L_{K+1} - \log L_K}{\log L_2 - \log L_1}, \tag{5.3}$$

which yields a number between 0 and 1, where a small value indicates that the K-class model can be used as the first split, while a larger value indicates that the tree might improve with an additional class at the root of the tree. Note that instead of an increase in log-likelihood, in Equation 5.3 one may use other measures of improvement in fit, such as the decrease of the BIC or the AIC.

The procedure described above concerns LC analysis with cross-sectional data. However, if the recorded responses are repeated/longitudinal measurements of the same variable, the procedure can also be carried out with a Latent Class Growth (LCG) model. Such a model is very similar to a standard LC model, except that the class-specific conditional response probabilities are now restricted using a regression model containing time variables as predictors (typically a polynomial). By using a similar stepwise estimation algorithm as described above, one can also construct a tree version of LCG model, which we called a Latent Class Growth Tree (LCGT). This was described in more detail in Chapter 4.

5.2.2 Step 2: Quantifying the Classification Errors of every split of the LCT

To relate the external variables to latent class membership, in the second step one needs to assign the respondents to classes, which is usually done with the posterior membership probabilities. The two most popular assignment methods are modal and proportional assignment. Modal assignment consists of assigning a respondent to the class with the largest estimated posterior membership probability. This is also known as hard partitioning and can be conceptualized as a respondent having a weight of one for the class with the largest estimated posterior membership probability and zero for the other classes. Proportional assignment, also known as soft partitioning, implies that the weights of one respondent are equal the posterior membership probability of the corresponding class.

Irrespective of the assignment method used, the true (X) and assigned (W) class membership scores will differ. That is, classification errors are inevitable. As proportional assignment is what is used to build a LCT, this is also the method we will use for the classification itself and for the determination of the classification errors at each split.

After class assignment, the assignment variable W will require a correction for classification errors in the third step. Hence, the amount of error in it must first be calculated (Bolck et al., 2004). The amount of classification errors can be expressed as the probability of an assigned class membership s conditional on the true class membership k (Vermunt, 2010). For every split of the LCT, this can be assessed as follows:

$$P(W = s|X_c = k, X_p) = \frac{\frac{1}{N_p}\sum_{i=1}^{N} w_{p,i} P(X_c = k|\mathbf{y}_i, X_p) P(W = s|\mathbf{y}_i, X_p)}{P(X_c = k|X_p)}. \tag{5.4}$$

The main modification compared to the equation in the case of a standard LC model is that we have to account for the contribution of every individual at the parent node concerned, which is achieved with the weight $w_{p,i}$ indicating the person i's prevalence in the node concerned. The total 'sample' size, which is denoted as (N_p, is obtained as the sum of the $w_{p,i}$. Note that most of the terms are conditional on the parent node concerned.

5.2.3 Step 3: Relating class membership with external variables

After the tree has been built in the first step and the classification and their errors have been assessed in the second step, the third and final step consists of relating the class memberships and some external variables while correcting for the classification errors. Two different types of relation between class memberships and external variables can be assessed. The goal can either be to investigate how the responses to a certain variable differ across classes (e.g., is there a difference in age between the classes), or the goal can be to investigate to what extent a variable predicts class membership (e.g., does age influence the probability of belonging to a certain class). The first variant, in which the one compares the distribution on an external variable Z_i across latent classes, is defined as follows:

$$P(W = s, Z_i|X_p) = \sum_{k=1}^{K} P(X_c = k|X_p) f(Z_i|X_c = k, X_p) P(W = s|X_c = k, X_p), \tag{5.5}$$

while the second option, in which the external variables are covariates predicting class membership, is defined as follows:

$$P(W = s|Z_i, X_p) = \sum_{k=1}^{K} P(X_c = k|Z_i, X_p) P(W = s|X_c = k, X_p). \tag{5.6}$$

As pointed out by Vermunt (2010) and Bakk et al. (2013), both Equation (5.5) and (5.6) are basically LC models, in which the classification errors

$P(W = s|X_c = k, X_p)$ can be fixed to their values obtained from the second step. These model can be estimated either by maximum likelihood estimation (Vermunt, 2010) or by a specific type of weighted analysis, also referred to as the BCH-approach (Bolck et al., 2004). The ML option is the best option when the external variables serve as covariate of class membership, while the BCH approach is the more robust option when the external variables are distal outcomes (Bakk & Vermunt, 2016).

To build a LCT and apply the three-step method, we have developed an R-package (R Core Team, 2016), called LCTpackage, which uses the Latent GOLD 5.1 program (Vermunt & Magidson, 2013) for the actual parameter estimation at step one and step three. Apart from dealing with logistics of performing the many separate steps required to build a tree and perform the subsequent step-three analyses, the LCTpackage provides various visual representations of the constructed tree, including one showing the step-three information about the external variables at each of the nodes. In Appendix C, we provide R-code which illustrates, based on the empirical examples in this paper, how one can perform bias-adjusted three-step LCT modeling.

5.3 Empirical examples

5.3.1 Example 1: Social Capital

The data set in this first example comes from a study by Owen and Videras (2008) and contains a large number of respondents and indicators, corresponding to applications for which LCTs are most suited. Owen and Videras (2008) used the information from 14.527 respondents of several samples of the General Social Survey to construct "a typology of social capital that accounts for the different incentives that networks provide." The data set contains sixteen dichotomous variables indicating whether respondents participate in specific types of voluntary organizations (the organizations are listed in the legend of Figure 5.2) and two variables indicating whether respondents agree with the statements "other people are fair" and "other people can be trusted". In this example these variables are used to build a LCT for this data set and the three-step procedure for LCTs is used to assess class differences in several demographic variables, to with age and gender. For this example we estimate the step-three model of every split with the BCH-approach (Vermunt, 2010), as this is the preferred option for continuous distal outcomes (Bakk & Vermunt, 2016).

Table 5.1: Log-likelihood, number of parameters, BIC, AIC, and relative improvement of the fit statistics of a traditional LC model with 1 to 9 classes.

	$\log L$	P	BIC	AIC	$RI_{\log L}$	RI_{BIC}	RI_{AIC}
1	-94204	3	188581	188444			
2	-89510	7	179376	179095			
3	-88501	11	177539	177115	0.215	0.199	0.212
4	-88117	15	176952	176383	0.082	0.064	0.078
5	-87826	19	176553	175840	0.062	0.043	0.058
6	-87619	23	176321	175464	0.044	0.025	0.040
7	-87425	27	176114	175113	0.041	0.022	0.038
8	-87322	31	176090	174945	0.022	0.003	0.018
9	-87234	35	176098	174808	0.019	-0.001	0.015

To decide on the number of classes at the root of the tree, multiple standard LC models with increasing number of classes are estimated. The fit statistics and the relative improvement of the fit statistics are shown in Table 5.1. The relative fit improvement is about 20% when expanding a model from 2 to 3 classes, compared to the improvement in fit when expanding from 1 to 2 classes. Adding more classes improves the fit only marginally and thus a root size of three classes is used. The final LCT is shown in Figure 5.1, with the class sizes displayed for every node of the tree. For every final node it holds that, according to the BIC, a 1-class model is preferred to a 2-class model.

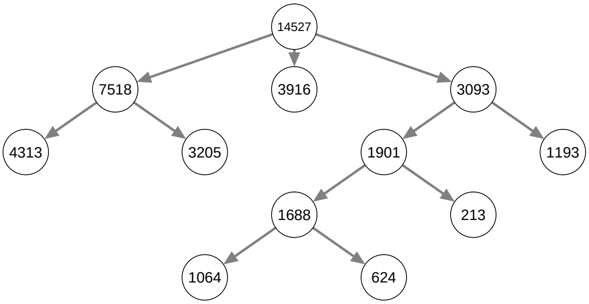

Figure 5.1: Layout of a a LCT with a root of three classes on social capital.

To interpret the tree, the profile plots of every split, as shown in Figure 5.2, can be investigated. The first split shows three classes, of which the first is has a low probability on all variables, the second displays a low probability on participation in all voluntary organizations and very high probabilities on the variables fair and trust, while the third class displays relative high probabilities on participation in the voluntary organizations and rather high

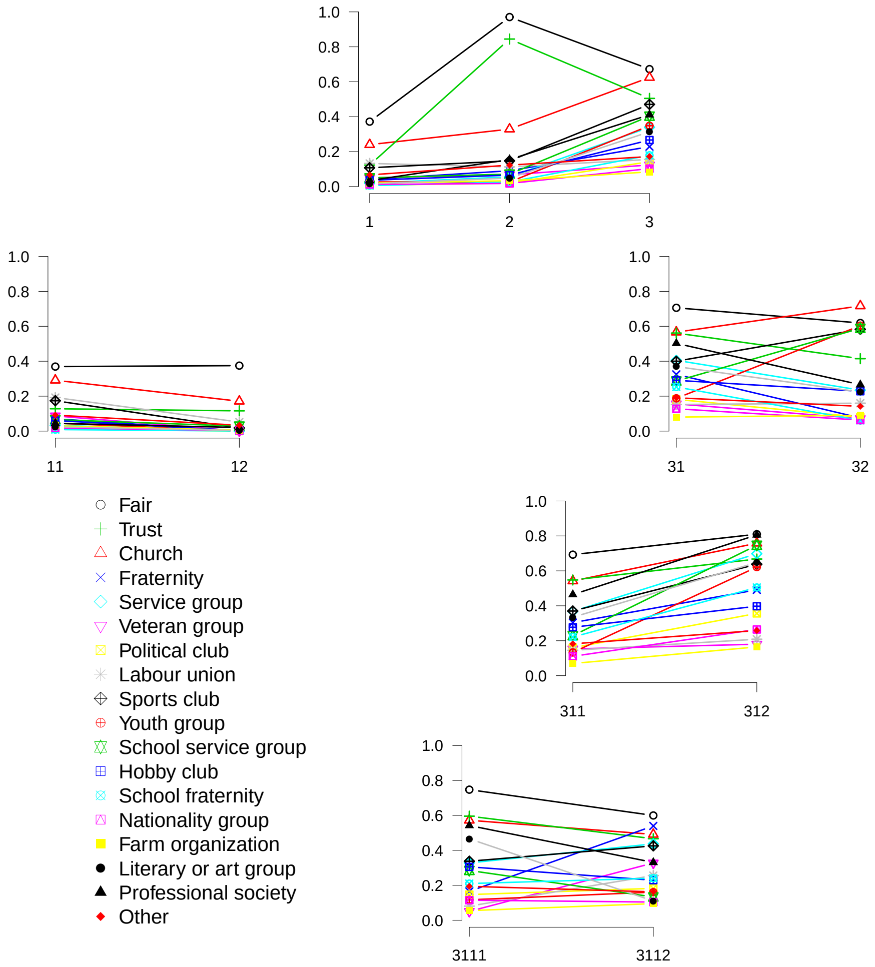

Figure 5.2: Profile plots of a LCT with a root of three classes on social capital.

probabilities for fair and trust. Subsequently, the first and third class are split further, while the second is not. The first class is split in a class with low and very low probabilities on all variables, while the third class is split in two classes with preferences for different voluntary organizations (e.g., a high probability for being part of a professional organization in class 31 versus a high probability for being part of a youth group in class 32). Subsequently class 31 is split further, in classes 311 and 312, which seem to differ mainly in participation in all voluntary organizations. The final split in classes 3111 and 3112 results in classes which differ again in preferences for different voluntary organizations (e.g, a high probability for being part of a literary or art group in class 3111 versus a high probability for being part of a fraternity in class 3112).

After building the tree, the three-step procedure is used to investigate the differences in the continuous variable age and the dichotomous variable gender given the class memberships. Between all classes of every split the mean age is compared, while for the variable gender a cross table is created and the percentage of (wo)men is compared between the classes. The results of the three-step method are visually displayed in Figure 5.3. From this figure we can conclude that after the first split the age is highest in class 2

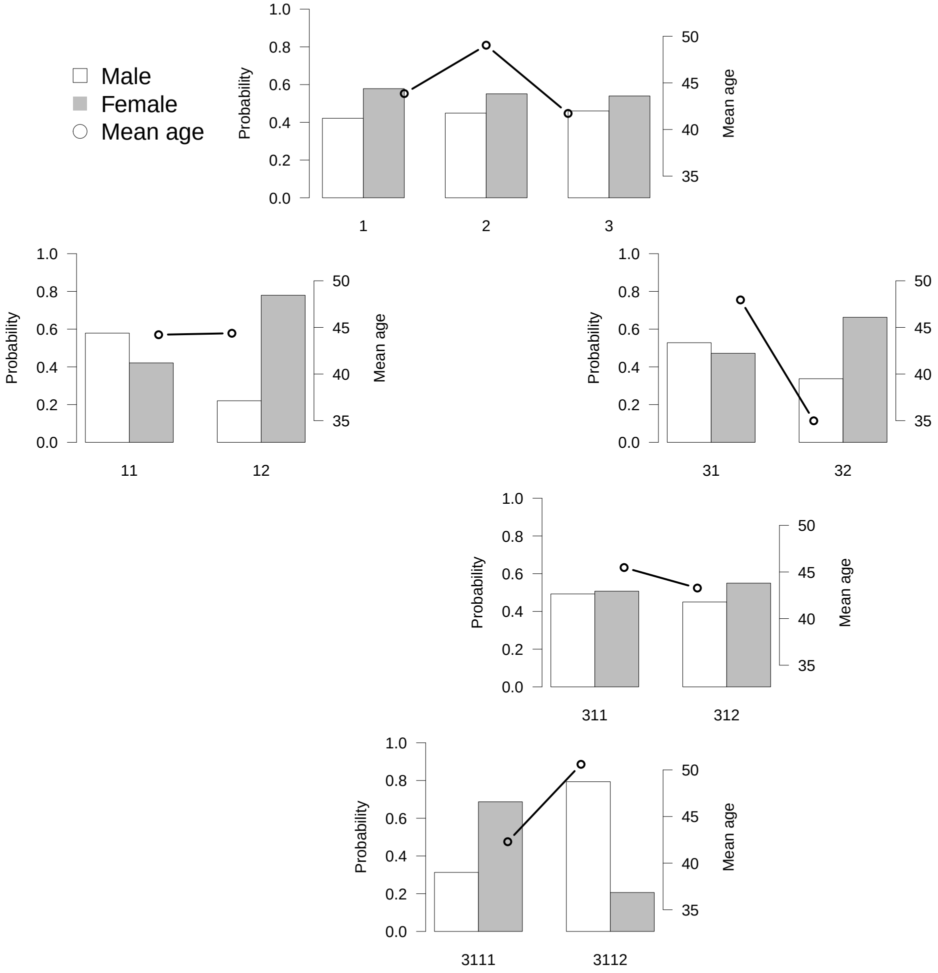

Figure 5.3: Results of the three-step procedure for gender and age on the LCT on social capital.

and lowest in class 3, while the percentage of (wo)men is about the same in every class, though still significantly different according to a Wald test ($W(2)$=11.690, p<0.05). After the split of class 1 there is no noticeable difference in age between classes 11 and 12, as can be seen in Figure 5.3 and this is also confirmed by a Wald test ($W(1)$=0.040, p=0.84). There is a significant difference in the percentage of (wo)men between classes 11 and 12 ($W(1)$=192.656, p<0.05). It seems that class 12, with very low probabilities on all variables, mainly consists of women, while class 11, with low probabilities on all variables, consists of more men. The split of class 3 results in two classes which differ both on average age ($W(1)$=258.988, p<0.05) and percentage of (wo)men ($W(1)$=46.090, p<0.05). The difference in age between these classes (and the direction of the difference) could be explained by the fact that class 31 contains more respondents that are part of a professional organization, while class 32 contains more respondents that are part of a youth group and the latter are a lot younger than the former. The difference in the proportion of men and women is not that large in class 31 (53% men and 47% women), while this difference is quite profound in class 32 (34% men and 66% women). The next split in classes 311 and 312 does not result in any significant differences on age ($W(1)$=2.090, p=0.15) and percentage of (wo)men ($W(1)$=0.746, p<0.39), while the final split in classes 3111 and 3112 results in differences in both age ($W(1)$=116.411, p<0.05) and percentage of (wo)men ($W(1)$=42.934, p<0.05).

5.3.2 Example 2: Mood Regulation

The second data set stems from a momentary assessment study by Crayen et al. (2012). It contains 8 mood assessments per day during a period of one week among 164 respondents (88 women and 76 men, with a mean age of 23.7, SD = 3.31). Respondents answered a small number of questions on a handheld device at pseudo-random signals during their waking hours. The delay between adjacent signals could vary between 60 and 180 minutes (M [SD] = 100.24[20.36] minutes, min = 62 minutes, max = 173 minutes). Responses had to be made within a 30-minute time window after the signal, and were otherwise counted as missing. On average, the 164 participants responded to 51 (of 56) signals (M [SD] = 51.07 [6.05] signals, min = 19 signals, max = 56 signals). In total, there were 8374 non-missing measurements.

At each measurement occasion, participants rated their momentary mood on an adapted short version of the Multidimensional Mood Questionnaire

(MMQ). Instead of the original monopolar mood items, a shorter bipolar version was used to fit the need for brief scales. Four items assessed pleasant-unpleasant mood (happy-unhappy, content-discontent, good-bad, and well-unwell). Participants rated how they momentarily felt on a 4-point bipolar intensity scales (e.g., very unwell, rather unwell, rather well, very well). For the current analysis, we focus on the item well-unwell. Preliminary analysis of the response-category frequencies showed that the lowest category (i.e., very unwell) was only chosen in approximately 1% of all occasions. Therefore the two lower categories were collapsed together into one unwell category. The following LCGT model is based on the recoded item with three categories (conform Crayen et al. (2012)). Subsequently the three-step procedure is applied to this LCGT. For the subsequent bias-adjusted three-step tree procedure, three personality traits (neuroticism, extraversion and conscientiousness) are used to predict latent class membership. These traits were assessed with the German NEO-FFI (Borkenau & Ostendorf, 2008) before the momentary assessment study started. The score of each trait is a mean of twelve items per dimension, ranging from 0 to 4. For this example, we estimate the step-three model of every split with maximum likelihood estimation (Vermunt, 2010), as this is the preferred option when the external variables as used as covariates predicting class membership (Bakk & Vermunt, 2016).

For the analysis, we used a LCG model based on an ordinal logit model. The time variable was the time during the day, meaning that we model the mood change during the day. There was a substantial difference between a tree based on a second- or a third-degree polynomial, which indicates that developments are better described by cubic growth curves than quadratic growth curves (see also the trajectory plots in Figure 5.5). Because there was no substantial difference between a tree based on a third- or a fourth-degree polynomial, a third-degree polynomial was used. Based on the relative improvement of the log-likelihood, BIC, and AIC (Table 5.2), it seems sensible to increase the number of classes at the root of the tree to three.

The layout and size of the LCGT with three root classes are presented in Figure 5.4 and its growth curve plots in Figure 5.5. The growth plots show that at the root of the tree, the three different classes all improve their mood during the day. They differ in their overall mood level, with class 3 having the lowest and class 2 having the highest overall score. Moreover, class 1 seems to be more consistently increasing than the other two classes. These three classes can be split further. Class 1 splits into two classes with both an

Table 5.2: Log-likelihood, number of parameters, BIC, AIC, and relative improvement of the fit statistics of a traditional LC growth model with 1 to 9 classes.

	$\log L$	P	BIC	AIC	$RI_{\log L}$	RI_{BIC}	RI_{AIC}
1	-7199	4	14424	14408			
2	-6741	9	13538	13504			
3	-6578	14	13244	13191	0.355	0.333	0.347
4	-6516	19	13149	13077	0.137	0.107	0.126
5	-6471	24	13091	13001	0.097	0.065	0.085
6	-6443	29	13064	12956	0.062	0.030	0.050
7	-6424	34	13058	12931	0.040	0.007	0.028
8	-6415	39	13069	12923	0.021	-0.013	0.008
9	-6404	44	13078	12914	0.024	-0.010	0.011

average score around one, class 11 just above and class 12 just below. Moreover, the increase in class 11 is larger than in class 12. The split of class 2 results in class 21 consisting of respondents with a very good mood in the morning, a rapid decrease until mid-day, and a subsequent increase. In general, the mean score of class 21 is high relative to the other classes. Class 22 starts with an average mean score and subsequently only increases. The splitting of class 3 results in two classes with a below average mood. Both classes increase, class 31 mainly in the beginning and class 32 mainly at the end of the day.

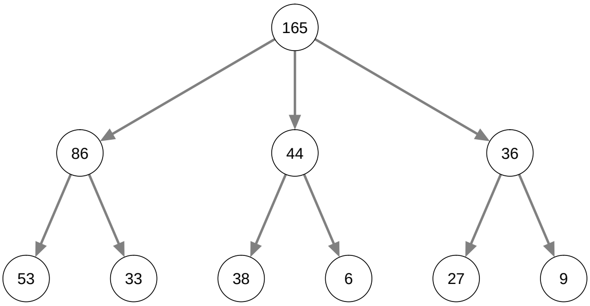

Figure 5.4: Layout of a LCGT with a root of three classes on mood regulation during the day.

After building the tree, the three-step procedure is used to investigate the relation of the three personality traits (neuroticism, extraversion and conscientiousness) with latent class membership. It is investigated to what extent each of the personality traits can predict latent class membership, while controlling for the other traits. The results of this three-step procedure are depicted in two separate figures, as the root of the tree splits into three classes

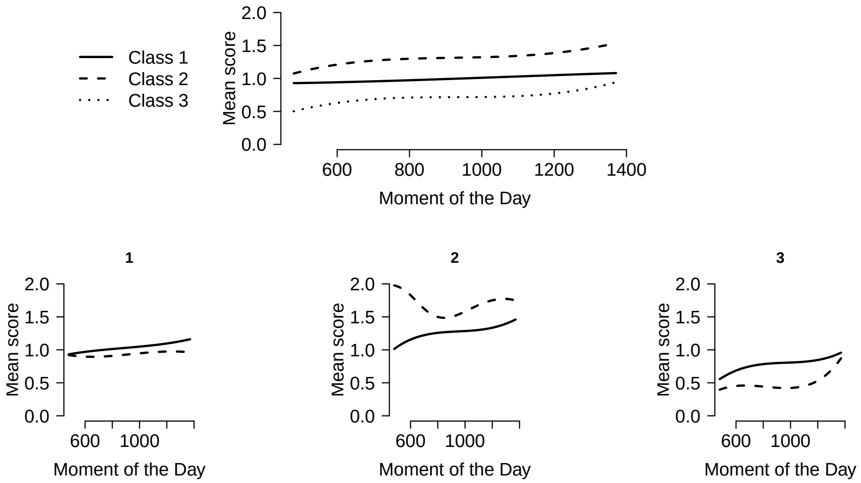

Figure 5.5: Profile plots of a LCGT with a root of three classes on mood regulation during the day.

and is more complex than the subsequent splits of the tree. In Figure 5.6 the results of the tree-step procedure on the first split are displayed for every variable separately. Each line indicates the probability of belonging to a certain class given the score of one of the personality traits. Note that the probability of belonging to a certain class depends on the combined score of the personality traits. Therefore, the displayed probability for each trait is conditional on the average of the other two traits.

The first graph of Figure 5.6 shows that a person with a low score on neuroticism has a relatively high probability of belonging to class 1. However, this probability decreases when neuroticism increases and when a person has a score on neuroticism above three this person is most likely to belong to class 3. Hence, a very neurotic person is likely to display a low overall mood level, while less neurotic persons are most likely to display a mood level that is neither very high nor very low. The second graph of Figure 5.6 shows that a person with a low score on extraversion has a high probability of belonging to class 1, but when extraversion decreases, so does the probability of belonging to class 1. Respondents with a score of 3.7 or higher on extraversion most likely belong to class 2. Hence, a very extravert person likely has a high overall positive mood level, while less extravert persons are most likely to display a positive mood level that is neither very high nor very low. The last graph of Figure 5.6 shows that a person with a low score on conscientiousness is most likely to belong to class 3. When a person score on conscientiousness is above 1.6, this person is more likely to belong to class 1. This indicates that persons

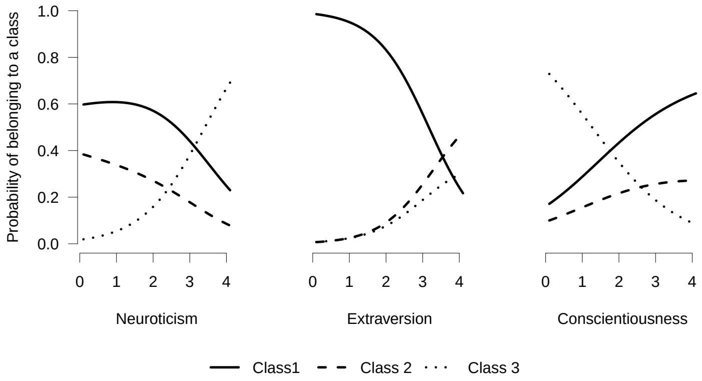

Figure 5.6: Results of the three-step procedure for the three personality traits on the root of the LCGT on mood regulation during the day.

with a low conscientiousness are most likely to display a non-positive overall mood level, while persons with a high conscientiousness are most likely to display an average mood level.

Figure 5.7 shows the results of the tree-step procedure on each of the three splits at the second level of the LCGT on mood regulation. Each graph shows the results for one split and every line indicates the probability of belonging to the first and largest class of the split corresponding to the personality trait in question (again conditional on an average score of the other two personality traits). The probability of belonging to the second class is not displayed, but when there are only two classes this is by definition the complement of the probability of belonging to the first class. Note that these results are conditional on being in class 1, 2, or 3.

The first graph of Figure 5.7 shows that the probability of belonging to class 11 increases mainly with a low score on conscientiousness and/or a high score on extraversion. The effect of neuroticism is less strong, but a higher score does indicate a higher probability of belonging to class 11. Hence, low conscientiousness, high extraversion, or high neuroticism indicate a higher probability that respondents' mood is in general slightly more positive. The second graph of Figure 5.7 shows that class membership is not really influenced by different scores of the personality traits, but only when extraversion is very high. Hence, the three personality trait are not good predictors for whether a respondent of class 2 has a somewhat continuously rising mood,

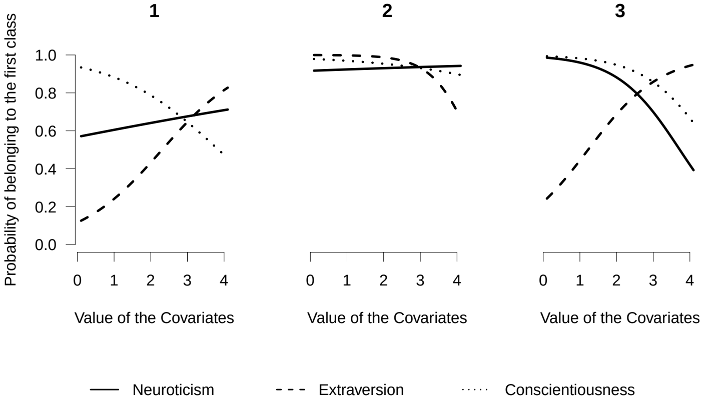

Figure 5.7: Results of the three-step procedure for the three personality traits on the second level of the LCGT on mood regulation during the day.

or a higher, but more fluctuating mood. The third graph of Figure 5.7 shows that a person with a low neuroticism, low conscientiousness, and/or high extraversion is most likely to be a member of class 31, while a person with a high neuroticism, high conscientiousness or low extraversion is most likely to be a member of class 32. Hence, a person with a high score on neuroticism or conscientiousness or a low score on extraversion is more likely to have a more negative overall mood than a person with a low score on neuroticism or conscientiousness or a high score on extraversion.

5.4 Discussion

LC and LCG models are used by researchers to identify (unobserved) subpopulations within their data. Because the number of latent classes retrieved is often large, the interpretation of the classes can become difficult. LCT and LCGT modeling has been developed to deal with this problem. However, assessing and interpreting the classes in LC and LCG models is usually the first part of an analysis with latent classes. Quite often researchers are interested in the relation between the classes and some external variables. This is commonly done by performing a second step in which respondents are assigned to the estimated classes and a third step in which the relationship of interest is studied using the assigned classes, where in the latter step one may also take the classification errors into account to prevent possible bias in the estimates. In this paper, we have shown how to transform the bias-adjusted three-step LC procedure to be applicable also in the context of LCT modeling.

The bias-adjusted three-step approach for LCT modeling has been illustrated with two empirical examples, one in which external variables are treated as distal outcomes of class membership and one in which external variables are used as predictors of class membership. The three-step approach as presented in this paper yields results per split of the LCT. A alternative could be to decide on the final classes of the LCT, and subsequently apply the three-step procedure to these end node classes simultaneously. This comes down to applying the original three-step approach, but neglecting that a LCT is built with sequential splits. Since these sequential splits are one of the main benefits of LCTs which facilitate the interpretation of the classes, the approach chosen here makes full use of the structure of a LCT. Another alteration could be to use modal assignment in step two instead of proportional assignment, with implies that one will have less classification errors. However, we do not expect this will matter very much since in the third step one takes into account the classification errors introduced by the classification method used (Bakk, Oberski, & Vermunt, 2014).

The bias-adjusted three-step method has become quite popular among applied researchers, but the basis of this method, the LC and LCG models, are not easy at all for applied researchers(Van de Schoot et al., 2017). The tree approach facilitates the use of these models, which can lead to more interpretable classes. With the addition of the bias-adjusted three-step method for LCTs and LCGTs, these classes can now also be related to external variables.

Chapter 6

Discussion and Conclusion

6.1 Short summary

This thesis introduced Latent Class Tree (LCT) analysis as an alternative for Latent Class (LC) analysis in situations where the LCs are difficult to interpret and/or the fit measures improve up to a large number of classes. This stepwise approach allows to use substantive information to decide whether or not to expand the number of classes.

Chapter 2 introduced the LCT procedure, which involves building a binary LCT based on 2-class LC models that are split repeatedly to impose a hierarchical structure on the classes. An empirical example on social capital was used to demonstrate the LCT analysis.

In Chapter 3, a hierarchical structure of classes was also imposed, but without the restriction of only binary splits. Because the first split of the tree picks up the most dominant associations in the data, we proposed a measure of relative improvement of fit to decide whether the number LCs at this first split of the data should be increased. This approach was illustrated again with an empirical example on social capital. Moreover, the LCT procedure was also applied to data from a cross-national study using a set of ranking items on (post-)materialism, which illustrated how the LCT procedure can accommodate for other types of LC models.

The tree procedure can also be applied to longitudinal data. In Chapter 4 the LCT procedure was extended to construct Latent Class Growth Trees (LCGT). A large number of classes is quite common with longitudinal data and the tree approach can be very suitable for assessing classes with different developmental patterns. LCGTs were illustrated in Chapter 4 with empirical examples on drugs use during adolescence and mood regulation during the day assessed using experience sampling.

Chapter 5 dealt with relating the classes of a LCT or LCGT to external variables, either covariates or distal outcomes. It was shown how to expand the three-step LC procedure to be applicable to each of the splits of the tree. This approach was illustrated with empirical examples using the data sets on social capital and mood regulation.

6.2 Future research

Besides the options of the LCT procedure developed in the chapters of this thesis, there are various other issues that could be important to consider when building a LCT. Below some of these issues, their implications, and possible extensions of the LCT methodology are discussed.

6.2.1 Statistics for model building

Throughout this thesis the decision to accept a split of a LCT has been based on the BIC because it is the most popular model selection measure among researchers using LC models, However, there are several other information criteria that can be used to decide on a split of a LCT. By using, for instance, the AIC (Akaike, 1974) or the SABIC (Sclove, 1987), the decision to accept a split is more likely, as these criteria are more lenient than the BIC. Note, that this will in general imply more splits and thus larger branches, but not necessarily more branches. It should be noted that the measure of relative fit improvement – used to decide about the number of classes at the first split – is less affected by the chosen criterion as the penalty terms for the number of parameters cancel each other out to a large extent.

Other criteria that take more than just the fit into account can be used. For instance, the entropy is a measure for the similarity between classes, based on the posterior class membership probabilities. The Classification Likelihood Criterion, the Integrated Classification Likelihood BIC, and the Approximate Weight of Evidence are all based on a combination of the fit of a model and its entropy (Vermunt & Magidson, 2013). When the goal is to classify respondents based on the model, possibly for relating the classes to external variables using the three-step method, such measures can become more important. Besides existing criteria also new criteria could be developed for LCTs. For example, while in this thesis the penalty term of the BIC used within every LCT was based on the total sample size, this could be modified to the sample size of the class at hand.

Another option for deciding whether to accept a split is the use of likelihood-ratio tests. A standard likelihood-ratio test can not be used for comparing models with different number of classes, as it is not chi-square distributed (McLachlan & Peel, 2004). However, the Lo-Mendell-Rubin likelihood ratio test (Lo, Mendell, & Rubin, 2001) and the bootstrap likelihood ratio test (McLachlan & Peel, 2004) are often used to choose between a $K-1$ classes and a K classes model. These tests provide a p-value and thus a more formal test to make a decision. However, the Lo-Mendell-Rubin likelihood-ratio test is somewhat controversial (Jeffries, 2003), while the bootstrap likelihood-ratio test is computationally intensive. Therefore, throughout this thesis the LCT procedure has been illustrated with the most commonly used information criterion, the BIC.

6.2.2 The splitting method of the divisive algorithm

Another issue relevant for LCTs is the occurrence of classification errors, which are basically always present in any LC model. The LC model is a probabilistic model providing K class membership probabilities for every response pattern. Suppose, for example, 100 respondents with the same response pattern are to be allocated in a 2-class model and the posterior class membership probabilities for the two classes equal 0.90 and 0.10. The classification errors depend (among others) on the choice of class assignment rule. When using modal assignment all 100 respondents are assigned to the first class, and 10 of them will thus be assigned to the wrong class. With proportional assignment, we use a 0.90-0.10 split of every person, which gives a total of 100*(0.90*0.10+0.10*0.90)=18 classification errors.

In LCT modeling, class assignments are not only used after model estimation, but also during model estimation to obtain the splits. Proportional assignment was used for this purpose in the LCT models described in this thesis, which is in agreement with the divisive LC algorithm proposed by Van der Palm et al. (2016). It implies constructing new data sets with case weights equal to the posterior class membership probability for the class and respondent concerned. However, a LCT could also be constructed based on an alternative assignment method, such as modal assignment. With modal assignment, respondents are assigned only to the class with the highest (modal) posterior probability. Hence, after the first split, subsequent splits will be based on subsets of the original sample. This is contrary to proportional assignment, where it is possible that some respondents receive a low weight, but these will not become exactly zero. Besides modal and proportional assignment, which are the most common assignment methods, it is also possible to use random assignment of individuals to classes based on their posterior class membership probabilities. This is basically a stochastic version of the proportional assignment rule. Under random and proportional assignment the expected number of misclassifications is the same, but random assignment ensures that not parts of one respondent are assigned to different classes, as happens with proportional assignment. Additionally there is also the option to assign respondents to a class if the posterior membership probability is larger than a certain threshold, which is more related to modal assignment.

6.2.3 Merging classes

Because the posterior probabilities are multiplied throughout the tree to construct new weights, the classification errors will contaminate further classes and splits. This can be handled in part by avoiding too small splits, as discussed in Chapter 2. However, because LCT analysis is most applicable to data for which a large number of classes is retrieved, it is still warranted that within a large LCT similar classes could be retrieved. Hence, development of a procedure for merging similar LCs within a LCT could be a very useful topic for future research. Hennig (2010) already noted that for Gaussian mixture components the merging problem is very relevant in practice, but that its solution may be regarded as somewhat arbitrary. Therefore he does not advertise a single method that may be "optimal" in some sense, but suggests several alternatives approaches, and indicates the choice between them is always dependent on the aim of clustering. A similar approach could be followed when developing alternative methods for merging classes from a LCT. Note a similar thing applies to the relative improvement of fit measure introduced in Chapter 2, for which we did not provide a cut-off value since what is a large enough improvement depends on the research topic at hand.

6.2.4 Using resampling methods

The LCT procedure could also be extended by combining it with resampling methods. Note that a LCT is similar to a decision tree (Breiman et al., 1984). The main difference is that in decision trees splits are based on the association of a single response variable with a set of predictors, while in LCTs splits are based on the association among multiple response variables simultaneously. Resampling methods have become very popular for decision trees to improve classification and many of these methods could also be applied to or adapted for LCTs. For example, the bagging procedure proposed by Breiman (1996) - which involves combining classifications based on randomly generated training sets - may also be used in the context of LCTs. By building multiple LCTs based on different training sets, the stability of splits within a LCT can be studied. Another resampling method is cross-validation, which can help to assess the generalizability of a LCT. Additionally, an approach similar to random forest could be used to assess the importance of variables within a split.

6.2.5 Dealing with non-categorical data

The LCT models and empirical examples within this thesis have been based on categorical indicators. However, researchers also apply LC models with continuous or mixed indicators. LC analysis with continuous variables, which is also referred to as latent profile analysis, involves estimating the class-specific means and variances for the variables used in the analysis. Since continuous variables are more informative than categorical variables, in these applications one will often end up with a large number of classes. This implies the LCT procedure may also be very useful in such situations. In principle, the basics of the LCT approach can also be applied to continuous indicators to construct Latent Profile Trees.

6.3 Final remarks

Ultimately, the usefulness of the LCT method must be demonstrated by its frequent use by applied researchers. To facilitate the use, an R-package[*] is currently under development which automates the construction LCTs. This package calls the Latent GOLD program (Vermunt & Magidson, 2016) to recursively estimate LC models using appropriately weighted data sets, and builds the tree and the corresponding graphical displays.

Because the LCT procedure can be applied to basically any LC model, it can be used for very different types of research topics. Furthermore, depending of the researcher's interpretation of the splits as being substantially relevant yes or no, a different LCT can be obtained. However, there is in

[*] The developmental version can be found at `www.github.com/MattisvdBergh/LCT`

general not one particular clustering method which is correct in all situations (Hennig, 2015), and the same applies to LCTs that can yield different results depending on the research topic at hand. One of the strength of LCT modeling is that it offers more possibilities to substantiate different trees (and thus LC solutions). Such substantial reasoning can be more important than purely statistical reasoning aimed at improving model fit. This thesis shows that imposing a hierarchical structure on the discovered LCs yields a convenient method to facilitate the substantive interpretation of the differences between the classes. We hope that over the course of time the LCT method becomes a standard and frequently used alternative for LC analysis.

Appendix A

Supplemental Material Chapter 2

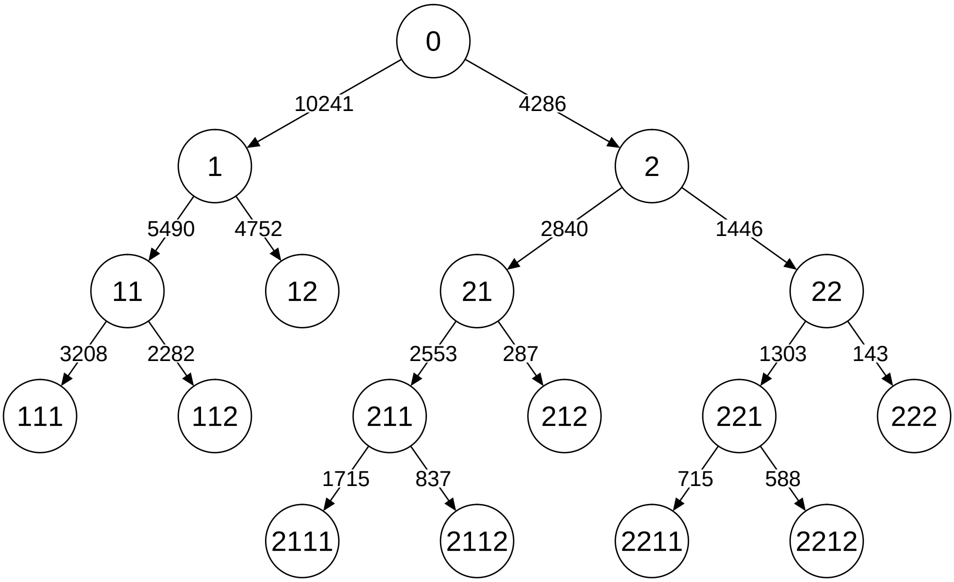

Figure A.1: Binary LCT based on the data of Owen & Videras (2009)

Table A.1: Conditional probabilities and class sizes of the binary LCT on Social Capital

	1	2	11	12	21	22	111	112	211	212	221	222	2111	2112	2211	2212
Fair	0.54	0.74	0.28	0.84	0.92	0.39	0.29	0.26	0.92	0.97	0.37	0.54	0.90	0.95	0.42	0.30
Trust	0.32	0.59	0.04	0.64	0.80	0.16	0.05	0.03	0.79	0.88	0.15	0.28	0.78	0.82	0.16	0.14
Frat	0.05	0.21	0.04	0.06	0.21	0.20	0.06	0.00	0.19	0.43	0.19	0.34	0.25	0.05	0.09	0.30
Serv	0.02	0.29	0.02	0.02	0.28	0.30	0.03	0.00	0.23	0.66	0.27	0.66	0.27	0.16	0.20	0.34
Vet	0.05	0.12	0.05	0.06	0.10	0.14	0.08	0.00	0.10	0.11	0.13	0.21	0.13	0.05	0.07	0.21
Polit	0.01	0.12	0.01	0.01	0.11	0.13	0.01	0.00	0.09	0.30	0.11	0.37	0.10	0.06	0.07	0.15
Union	0.13	0.15	0.13	0.12	0.13	0.18	0.19	0.05	0.13	0.14	0.18	0.23	0.13	0.11	0.18	0.18
Sport	0.10	0.41	0.11	0.10	0.36	0.51	0.17	0.01	0.34	0.54	0.48	0.77	0.29	0.46	0.55	0.39
Youth	0.02	0.27	0.03	0.01	0.19	0.43	0.05	0.00	0.17	0.42	0.39	0.77	0.06	0.40	0.62	0.12
School	0.05	0.33	0.05	0.04	0.28	0.44	0.07	0.03	0.24	0.57	0.40	0.82	0.13	0.48	0.58	0.18
Hobby	0.04	0.23	0.03	0.04	0.21	0.28	0.06	0.00	0.20	0.31	0.26	0.49	0.21	0.16	0.22	0.30
Sfrat	0.01	0.15	0.00	0.01	0.15	0.15	0.01	0.00	0.11	0.43	0.13	0.40	0.15	0.04	0.07	0.20
Nat	0.01	0.09	0.01	0.01	0.07	0.12	0.02	0.00	0.06	0.17	0.10	0.31	0.07	0.04	0.07	0.13
Farm	0.02	0.08	0.02	0.03	0.07	0.09	0.03	0.01	0.06	0.10	0.08	0.26	0.06	0.07	0.08	0.08
Lit	0.01	0.26	0.02	0.01	0.26	0.27	0.02	0.00	0.22	0.61	0.23	0.64	0.23	0.18	0.20	0.27
Prof	0.05	0.38	0.03	0.07	0.41	0.33	0.04	0.02	0.37	0.77	0.29	0.70	0.41	0.28	0.23	0.36
Church	0.25	0.59	0.24	0.26	0.57	0.63	0.29	0.16	0.55	0.69	0.60	0.88	0.49	0.68	0.69	0.48
Other	0.08	0.17	0.06	0.10	0.17	0.17	0.09	0.02	0.17	0.19	0.15	0.29	0.18	0.15	0.13	0.18

Table A.2: Conditional probabilities of the original 8-class LC model retrieved by Owen & Videras (2009)

	C1	C2	C3	C4	C5	C6	C7	C8
Fair	0.29	0.88	0.89	0.68	0.84	0.00	0.82	0.76
Trust	0.06	0.71	0.76	0.49	0.46	0.00	0.59	0.56
Frat	0.19	0.19	0.48	0.37	0.93	0.50	0.56	0.79
Serv	0.01	0.00	0.26	0.21	0.08	0.17	0.14	0.65
Vet	0.01	0.00	0.05	0.04	0.16	0.28	0.65	0.66
Polit	0.03	0.04	0.20	0.01	0.25	0.31	0.48	0.70
Union	0.08	0.11	0.30	0.21	0.08	0.42	0.68	0.64
Sport	0.01	0.02	0.03	0.08	0.10	0.06	0.06	0.16
Youth	0.02	0.04	0.22	0.12	0.13	0.23	0.14	0.38
School	0.01	0.01	0.32	0.02	0.12	0.19	0.08	0.58
Hobby	0.02	0.03	0.18	0.35	0.05	0.12	0.09	0.34
Sfrat	0.00	0.01	0.19	0.02	0.00	0.11	0.04	0.33
Nat	0.02	0.09	0.55	0.05	0.03	0.26	0.19	0.66
Farm	0.03	0.03	0.06	0.37	0.00	0.10	0.08	0.16
Lit	0.00	0.00	0.10	0.08	0.03	0.08	0.03	0.31
Prof	0.12	0.10	0.08	0.33	0.02	0.17	0.21	0.19
Church	0.01	0.01	0.08	0.04	0.02	0.08	0.02	0.21
Other	0.05	0.10	0.17	0.13	0.20	0.15	0.07	0.22
Class size	0.41	0.23	0.11	0.07	0.05	0.06	0.04	0.03

Appendix B

Supplemental Material Chapter 3

Table B.1: Conditional probabilities and class sizes of the LCT starting with a ternary split on Social Capital

	1	2	3	11	12	31	32	311	312	3111	3112
Fair	0.37	0.97	0.67	0.37	0.37	0.71	0.62	0.69	0.81	0.75	0.60
Trust	0.12	0.84	0.50	0.13	0.12	0.56	0.41	0.55	0.67	0.60	0.47
Frat	0.04	0.09	0.23	0.07	0.00	0.32	0.08	0.30	0.49	0.17	0.54
Serv	0.02	0.06	0.34	0.03	0.00	0.41	0.23	0.37	0.70	0.33	0.44
Vet	0.05	0.07	0.12	0.09	0.00	0.16	0.07	0.15	0.18	0.05	0.33
Polit	0.01	0.02	0.14	0.01	0.00	0.18	0.08	0.16	0.36	0.15	0.18
Union	0.13	0.11	0.16	0.19	0.05	0.15	0.16	0.15	0.21	0.08	0.26
Sport	0.11	0.15	0.47	0.17	0.02	0.40	0.58	0.37	0.64	0.34	0.43
Youth	0.03	0.03	0.35	0.05	0.00	0.19	0.60	0.13	0.62	0.12	0.16
School	0.05	0.07	0.40	0.07	0.03	0.29	0.59	0.23	0.75	0.28	0.13
Hobby	0.04	0.07	0.27	0.06	0.00	0.29	0.23	0.28	0.40	0.31	0.23
Sfrat	0.01	0.03	0.18	0.01	0.00	0.25	0.06	0.22	0.50	0.21	0.24
Nat	0.01	0.02	0.10	0.02	0.00	0.13	0.06	0.11	0.26	0.11	0.10
Farm	0.02	0.03	0.08	0.03	0.01	0.08	0.09	0.07	0.16	0.05	0.09
Lit	0.02	0.05	0.31	0.02	0.00	0.37	0.22	0.33	0.65	0.47	0.11
Prof	0.03	0.15	0.41	0.04	0.02	0.50	0.27	0.46	0.81	0.54	0.33
Church	0.24	0.33	0.62	0.29	0.17	0.57	0.72	0.54	0.76	0.57	0.49
Other	0.07	0.12	0.17	0.09	0.03	0.19	0.14	0.18	0.26	0.19	0.16
Class sizes	0.52	0.27	0.21	0.30	0.22	0.13	0.08	0.12	0.01	0.07	0.04

Appendix C

Supplemental Material Chapter 5

C.1 Code to build the LCT on social capital with the LCTpackage

While the LCTpackage will appear eventually on CRAN, it can currently only be obtained from the first authors Github page with the following code. Moreover, it requires installation of the Latent GOLD program (Vermunt & Magidson, 2013) for the main computations.

```
library(devtools)
install_github("MattisvdBergh/LCT")
library(LCTpackage)

# Filepath of the Latent GOLD 5.1 executable, e.g.:
LG = "C:/Users/Mattis/LatentGOLD5.1/lg51.exe"
```

The data set used in the first empirical example is part of the LCTpackage. The code below prepares the data and some arguments required for the LCT function.

```
data("SocialCapital")
itemNames = c("fair", "memchurh", "trust", "memfrat",
              "memserv", "memvet", "mempolit",
              "memunion", "memsport", "memyouth",
              "memschl", "memhobby", "memgreek",
              "memnat", "memfarm", "memlit",
              "memprof", "memother")
```

```
# Make the items factors, to be ordinal in the model
SocialCapital[itemNames] = sapply(
  SocialCapital[,itemNames],
  function(x){as.factor(x)})
```

The function below builds a LCT on social capital with a root of 3 classes. The Latent GOLD syntaxes for every split and its results are written in a newly created folder in the current working directory, called ResultsSC3.

```
Results.SC3 = LCT(Dataset = SocialCapital,
                  LG = LG,
                  resultsName = "SC3",
                  itemNames = itemNames,
                  nKeepVariables = 2,
                  namesKeepVariables = c("age", "sex"),
                  maxClassSplit1 = 3)
```

The function below applies the three-step method to the LCT on social capital with a root of 3 classes. The Latent GOLD syntaxes for every split and its results are written in a newly created folder in the current working directory, this time called exploreTreeSC3.

```
explTree.SC3 = exploreTree(resTree = Results.SC3,
                           sizeMlevels = c(2, 1),
                           dirTreeResults =
                             paste0(getwd(),
                                    "/ResultsSC3"),
                           ResultsFolder =
                             "exploreTreeSC3",
                           Covariates = c("sex", "age"),
                           mLevels = c("ordinal",
                                       "continuous"),
                           analysis = "dependent")
```

C.2 Code to build the LCGT of the example on mood regulation with the LCTpackage

The data set is again part of the LCTpackage. The code below prepares the data.

```r
data("MoodRegulation")
# Make the items factors, to be ordinal in the model
MoodRegulation[,"well"] =
  as.factor(MoodRegulation[,"well"])

# Recode male and female, to have the variable
# coded the same as in the social capital example
MoodRegulation[,"male"] =
  ifelse(MoodRegulation[,"male"]==1,
         1,
         ifelse(MoodRegulation[,"male"]==0,
                                        2,
                                        NA))
```

The function below builds a LCGT on mood regulation with a root of 3 classes. The Latent GOLD syntaxes for every split and its results are written in a newly created folder in the current working directory, called ResultsC3.

```r
Results.C3 = LCGT(Dataset = MoodRegulation,
                  LG = LG, dependent = "well",
                  independent = c("time_cont",
                                  "time_cont2",
                                  "time_cont3"),
                  caseid = "UserID",
                  levelsDependent = 3,
                  resultsName = "C3",
                  nKeepVariables = 3,
                  namesKeepVariables =
                    c("Neuroticism",
                      "Extraversion",
                      "Conscientiousness"),
                  maxClassSplit1 = 3)
```

The function below applies the three-step method to the LCGT on mood regulation with a root of 3 classes. The Latent GOLD syntaxes for every split and its results are written in a newly created folder in the current working directory, this time called exploreTreeC3.

```
explTree.C3 = exploreTree(resTree = Results.C3,
                          LG = LG,
                          sizeMlevels = rep(1, 3),
                          dirTreeResults =
                            paste0(getwd(),
                                   "/ResultsSC3"),
                          ResultsFolder =
                            "exploreTreeC3",
                          Covariates =
                            c("Neuroticism",
                              "Extraversion",
                              "Conscientiousness"),
                          mLevels =
                            rep("continuous", 3),
                          analysis = "covariates",
                          method = "ml")
```

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, *40*(2), 235–243. doi: 10.1509/jmkr.40.2.235.19225

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis*, *22*(4), 520–540. doi: 10.1093/pan/mpu003

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2016). Relating latent class membership to continuous distal outcomes: Improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 278–289. doi: 10.1080/10705511.2015.1049698

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*(1), 272–311. doi: 10.1177/0081175012470644

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 20–31. doi: 10.1080/10705511.2014.955104

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725. doi: 10.1109/34.865189

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27. doi: 10.1093/pan/mph001

Borkenau, P., & Ostendorf, F. (2008). *NEO-FFI : NEO-Fünf-Faktoren-Inventar nach Costa und McCrae, Manual*. Hogrefe: Göttingen.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013).

Structural equation model trees. *Psychological methods*, *18*(1), 71–86. doi: 10.1037/a0030001

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi: 10.1007/BF00058655

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. New York: Taylor & Francis.

Clogg, C. C. (1995). Latent class models. In *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York: Plenum Press.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., & Vermunt, J. K. (2012). Exploring dynamics in mood regulation—mixture latent Markov modeling of ambulatory assessment data. *Psychosomatic Medicine*, *74*(4), 366–376. doi: 10.1097/PSY.0b013e31825474cb

Croon, M. (1990). Latent class analysis with ordered latent classe. *British Journal of Mathematical and Statistical Psychology*, *43*(2), 171–192. doi: 10.1111/j.2044-8317.1990.tb00934.x

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*(401), 173–178. doi: 10.1080/01621459.1988.10478584

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*(4), 643–659. doi: 10.1007/s00180-007-0103-7

Ding, C., & He, X. (2002). Cluster merging and splitting in hierarchical clustering algorithms. In *IEEE International Conference on Data Mining, 2002. proceedings.* (pp. 139–146). doi: 10.1109/ICDM.2002.1183896

Dolan, C. V., & van der Maas, H. L. J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, *63*(3), 227–253. doi: 10.1007/BF02294853

Durlauf, S. N., & Fafchamps, M. (2004). Social capital. *mimeo*.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Hierarchical clustering. In *Cluster analysis* (pp. 71–110). John Wiley & Sons, Ltd. doi: 10.1002/9780470977811.ch4

Felipe, A., Miranda, P., & Pardo, L. (2015). Minimum phi-divergence estimation in constrained latent class models for binary data. *Psychometrika*, *80*(4), 1020–1042. doi: 10.1007/s11336-015-9450-4

Francis, B., Elliott, A., & Weldon, M. (2016). Smoothing group-based trajectory models through b-splines. *Journal of Developmental and Life-Course Criminology*, *2*(1), 113–133. doi: 10.1007/s40865-016-0025-6

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Berlin: Springer series in statistics.

Ghattas, B., Michel, P., & Boyer, L. (2017). Clustering nominal data using unsupervised binary decision trees: Comparisons with the state of the art methods. *Pattern Recognition*, *67*, 177–185. doi: 10.1016/j.patcog.2017.01.031

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231. doi: 10.1093/biomet/61.2.215

Haberman, S. J. (1979). *Analysis of qualitative data. vol. 2, new developments*. London: Academic Press.

Hadiwijaya, H., Klimstra, T. A., Vermunt, J. K., Branje, S. J. T., & Meeus, W. H. J. (2015). Parent–adolescent relationships: an adjusted person-centred approach. *European Journal of Developmental Psychology*, *12*(6), 728–739. doi: 10.1080/17405629.2015.1110519

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators. *Sociological Methods & Research*, *16*(3), 379–405. doi: 10.1177/0049124188016003002

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Sage Newbury Park.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage Publications, Inc.

Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification*, *4*(1), 3–34. doi: 10.1007/s11634-010-0058-3

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62. (Philosophical Aspects of Pattern Recognition) doi: 10.1016/j.patrec.2015.04.009

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(3), 309–369. doi: 10.1111/j.1467-9876.2012.01066.x

Inglehart, R. (1971). The silent revolution in europe: Intergenerational change in post-industrial societies. *American Political Science Review*, *65*(4), 991–1017. doi: 10.2307/1953494

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures (developmental psychology)*. New York: Basic Books.

Jansen, B. R., & van der Maas, H. L. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review*, *17*(3), 321–357. doi: 10.1006/drev.1997.0437

Jeffries, N. O. (2003). A note on 'testing the number of components in a normal mixture'. *Biometrika*, *90*(4), 991–994. doi: 10.1093/biomet/90.4.991

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research*, *29*(3), 374–393. doi: 10.1177/0049124101029003005

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*(1), 302–317. doi: 10.1111/j.1751-9004.2007.00054.x

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld, S. A. Star, & J. Clausen (Eds.), *Measurement and prediction* (Vol. 4, pp. 362–472). Princeton, NJ: Princeton university press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mill.

Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*(413), 96–107. doi: 10.1080/01621459.1991.10475008

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. doi: 10.1093/biomet/88.3.767

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*(4), 815–840.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*(1), 201–226. doi: 10.1146/annurev.psych.51.1.201

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology*, *31*(1), 223–264. doi: 10.1111/0081-1750.00096

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 175–198). Thousand Oakes: Sage Publications.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park: Sage Publications.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Moors, G., & Vermunt, J. (2007). Heterogeneity in post-materialist value priorities. evidence from a latent class discrete choice approach. *European Sociological Review*, *23*(5), 631–648. doi: 10.1093/esr/jcm027

Mulder, E., Vermunt, J., Brand, E., Bullens, R., & van Marle, H. (2012). Recidivism in subgroups of serious juvenile offenders: Different profiles, different risks? *Criminal Behaviour and Mental Health*, *22*(2), 122–135. doi: 10.1002/cbm.1819

Muthén, B. (2004). Latent variable analysis. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 345–68). Thousand Oakes: Sage Publications.

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*, *46*(1), 252–282. doi: 10.1177/0081175015581379

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2017). Power and type 1 error of local fit statistics in multilevel latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 216–229. doi: 10.1080/10705511.2016.1250639

Nagin, D. S. (2005). *Group-based modeling of development*. Harvard university press.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, *31*(3), 327–362. doi: 10.1111/j.1745-9125.1993.tb01133.x

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 92–105). Washington, DC: American Psychological Association.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. doi: 10.1080/10705510701575396

Oberski, D. L. (2016). Beyond the number of classes: separating substantive from non-substantive dependence in latent class analysis. *Advances in Data Analysis and Classification*, *10*(2), 171–182. doi: 10.1007/s11634-015

-0211-0

Oberski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, *7*(3), 267–279. doi: 10.1007/s11634-013-0146-2

Oberski, D. L., Vermunt, J., & Moors, G. (2015). Evaluating measurement invariance in categorical data latent variable models with the EPC-interest. *Political Analysis*, *23*(4), 550–563. doi: 10.1093/pan/mpv020

Oser, J., Hooghe, M., & Marien, S. (2013). Is online participation distinct from offline participation? a latent class analysis of participation types and their stratification. *Political Research Quarterly*, *66*(1), 91–101. doi: 10.1177/1065912912436695

Owen, A. L., & Videras, J. (2008). Reconsidering social capital: A latent class approach. *Empirical Economics*, *37*(3), 555–582. doi: 10.1007/s00181-008-0246-6

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. doi: 10.1177/014662169001400305

Santos, L. M., Amorim, L. D. A., Santos, D. N., & Barreto, M. L. (2015). Measuring the level of social support using latent class analysis. *Social Science Research*, *50*, 139–146. doi: 10.1016/j.ssresearch.2014.11.009

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., … Miles, A. (2013). A new model of social class? Findings from the BBC's great British class survey experiment. *Sociology*, *47*(2), 219–250. doi: 10.1177/0038038513481128

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343. doi: 10.1007/BF02294360

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520. doi: 10.1016/0010-0285(76)90016-5

Spycher, B. D., Silverman, M., Brooke, A. M., Minder, C. E., & Kuehni, C. E.

(2008). Distinguishing phenotypes of childhood wheeze and cough using latent class analysis. *European Respiratory Journal*, *31*(5), 974–981. doi: 10.1183/09031936.00153507

Studer, J., Baggio, S., Mohler-Kuo, M., Simon, O., Daeppen, J.-B., & Gmel, G. (2016). Latent class analysis of gambling activities in a sample of young swiss men: Association with gambling problems, substance use outcomes, personality traits and coping strategies. *Journal of Gambling Studies*, *32*(2), 421–440. doi: 10.1007/s10899-015-9547-9

Sullivan, P. F., Kessler, R. C., & Kendler, K. S. (1998). Latent class analysis of lifetime depressive symptoms in the national comorbidity survey. *American Journal of Psychiatry*, *155*, 1398–1406. doi: 10.1176/ajp.155.10.1398

Thullen, M. J., Taliaferro, L. A., & Muehlenkamp, J. J. (2016). Suicide ideation and attempts among adolescents engaged in risk behaviors: a latent class analysis. *Journal of Research on Adolescence*, *26*(3), 587–594. doi: 10.1111/jora.12199

Van den Bergh, M., Schmittmann, V. D., & Vermunt, J. K. (2017). Building latent class trees, with an application to a study of social capital. *Methodology*, *13*, 13–22. doi: 10.1027/1614-2241/a000128

Van der Heijden, P. G. M., Dessens, J., & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the em algorithm. *Journal of Educational and Behavioral Statistics*, *21*(3), 215–229. doi: 10.3102/10769986021003215

Van der Palm, D. W., van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification*, *33*(1), 52–72. doi: 10.1007/s00357-016-9195-5

Van de Schoot, R., Sijbrandij, M., Winter, S. D., Depaoli, S., & Vermunt, J. K. (2017). The grolts-checklist: Guidelines for reporting on latent trajectory studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 451–467. doi: 10.1080/10705511.2016.1247646

Van Hulst, B. M., de Zeeuw, P., & Durston, S. (2015). Distinct neuropsychological profiles within ADHD: A latent class analysis of cognitive control, reward sensitivity and timing. *Psychological Medicine*, *45*(4), 735–745. doi: 10.1017/S0033291714001792

Van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2015). Assessing model fit in latent class analysis when asymptotics do not hold. *Methodology*, *11*(2), 65–79. doi: 10.1027/1614-2241/a000093

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*(1), 213–239. doi: 10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K. (2007). Growth models for categorial response variables: Standard, latent-class, and hybrid approaches. In K. van Montfort, H. Oud, & A. Satorra (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 139–158). Mahwah, NJ: Erlbaum.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469. doi: 10.1093/pan/mpq025

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. P. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge, NY: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2013). Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. *Belmont, Massachusetts: Statistical Innovations Inc*.

Vermunt, J. K., & Magidson, J. (2016). Upgrade manual for Latent GOLD 5.1. *Belmont, Massachusetts: Statistical Innovations Inc*.

Vidotto, D., Kaptein, M. C., & Vermunt, J. K. (2015). Multiple imputation of missing categorical data using latent class models: State of the art. *Psychological Test and Assessment Modeling*, *57*(4), 542–576.

Wedel, M., & DeSarbo, W. S. (1994). A review of recent developments in latent class regression models. In R. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352–388). Cambridge, MA: Blackwell.

Yamaguchi, K. (2000). Multinomial logit latent-class regression models: An analysis of the predictors of gender role attitudes among japanese women. *American Journal of Sociology*, *105*(6), 1702–1740. doi: 10.1086/210470

Yung, Y.-F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, *62*(3), 297–330. doi: 10.1007/BF02294554

Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *The Journal of Machine Learning Research*, *5*, 697–723.

Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, *10*(2), 141–168. doi: 10.1007/s10618-005-0361-3

Summary

People differ, but some people are more alike to each other than to others. Within the social sciences, answer patterns based on variables of interest are used to cluster similar people. For example, based on variables on healthy and risky behavior (e.g., 'How often do you sport?' or 'How much do you smoke?') one could identify clusters of people showing qualitative different risk behavior. A popular method for this is Latent Class (LC) analysis which identifies unobserved homogenous subgroups or classes within a data set. A very important and sometimes difficult issue of LC analysis is the decision on how many classes should be used. This is usually decided by comparing models with a different number of classes on some measure that indicates how well the model describes/fits to the observed data. The number of classes is increased until the fit measure does not improve. Once the number of classes has been decided, the classes are interpreted based on the class specific probabilities or means.

This is a theoretically sound procedure, but in applied settings it is often problematic. For instance, when a LC analysis is applied to a large data set (with many respondents and/or many variables) there is frequently a large number of classes identified. Such a large number of classes is often more specific than intended and interpretation can become very hard. Moreover, the number of classes is only based on fit, as it is very hard to substantively compare models with different numbers of classes. Finally, the fact that different fit measures can indicate a different optimal number of classes does not help. In this theses LC models with different numbers of classes that are substantively related have been developed. These models are based on the so-called Latent Class Tree (LCT) procedure.

LCT modeling starts as a standard LC analysis by determining whether two classes fit better than one class. If the two class solution is preferred, the two classes are separated in a new data set for each of the classes and every respondent is proportionally assigned to each class. Subsequently a 1-class and a 2-class model are estimated for each of these classes. This procedure is repeatedly applied until only 1-class models are preferred. This results in

a hierarchical tree structure of latent classes. The main advantage of this approach is that classes will be substantially related. This allows the use of both statistical and substantive reasoning to decide on a number of classes. These binary LCTs are described in Chapter 2 and illustrated with an empirical data set on social capital.

The first split of a LCT is the most important split, as the rest of tree depends on this first result. Moreover, in some situations binary splits are too much a simplification and it is important to investigate this. Therefore in Chapter 3 a measure of the relative improvement in fit has been proposed to investigate models with different numbers of classes. Standard fit measures would result in standard LC analysis, but with the measure of relative improvement in fit it can be assessed whether it is preferable to increase the number of classes at the first split of the LCT. For subsequent splits there is more substantive information on the classes available to guide decisions on split sizes, while the relative fit improvement can also be used. The measure of relative improvement in fit is described in Chapter 3 and illustrated with several LCTs on social capital and (post-)materialism.

LC analysis is applied to cross-sectional data (assessed on one moment in time). For longitudinal data latent class growth curves are used. This identifies similar patterns over time, for example respondents with different mood patterns during the day. A large number of classes is quite common with longitudinal data and the tree approach can be very suitable for assessing classes with different patterns over time. Therefore in Chapter 4 the LCT approach has been expanded to also construct so-called Latent Class Growth Trees (LCGT). The LCGTs illustrated with empirical examples on mood regulation during the day and the probability of drugs use given a respondents age.

Assessing the latent classes is usually the first step of a study. Subsequently researchers often want to relate the classes to some external variables. For example, do some classes differ in the amount of men and women or can we predict class membership based on age. A procedure has been developed to compare the distributions of external variables among classes and to predict the class memberships based on external variables. This is applied to every split of a tree and therefore gives a clear overview on how the external variable is related to every class of the tree. This procedure is illustrated for both LCTs and LCGTs in Chapter 5.

The last chapter discusses possibilities for future research, such as resampling methods or specific fit measures. In principle the traditional LC analysis will be a first starting point for applied researchers, but with this work (and also the development of software to build LCTs) there is now the possibility in difficult situations to use the LCT method to have substantive information to decide on the optimal number of classes.

Samenvatting

Mensen verschillen van elkaar, maar sommige mensen lijken meer op elkaar dan andere mensen. Binnen de sociale wetenschappen worden mensen vaak gegroepeerd op basis van hun antwoordpatroon op een aantal variabelen. Bijvoorbeeld, bij vragen over gezondheid en risicofactoren (zoals: 'Hoe vaak sport je?' of 'Hoeveel rook je?') kunnen groepen mensen die kwalitatief verschillend risicogedrag vertonen worden geïdentificeerd. Een populaire methode hiervoor is Latente Klasse (LK) analyse, waarbij niet geobserveerde homogene groepen kunnen worden identificeerd in een dataset. Een belangrijke en soms moeilijke kwestie van LK analyse is het bepalen van het aantal klassen waarin mensen kunnen worden ingedeeld. Dit aantal wordt meestal bepaald door modellen met verschillende aantallen klassen met elkaar te vergelijken met een maat die aangeeft hoe goed het model de data beschrijft. Zolang een dergelijke fit wordt verbeterd, blijft het aantal klassen uitgebreid worden. Wanneer vervolgens het aantal klassen is bepaald, worden deze geïnterpreteerd aan de hand van klasse specifieke kansen of gemiddelden.

Er is theoretisch weinig mis met deze procedure, maar in de praktijk blijkt het vaak problematisch. Wanneer LK analyse wordt toegepast op een grote dataset (met veel personen en variabelen) dan wordt er regelmatig groot aantal klassen geïdentificeerd. Bij een groot aantal klassen zijn deze vaak specifieker dan gewenst en interpretatie is dan zeer lastig. Bovendien is het aantal klassen alleen gebaseerd op een fit maat, omdat het vaak heel moeilijk is om modellen met een verschillend aantal klassen inhoudelijk met elkaar te vergelijken. Verder helpt het niet dat verschillende fit maten kunnen leiden tot een ander aantal klassen. In dit proefschrift zijn LK modellen ontwikkelt waarbij de modellen met een verschillend aantal klassen duidelijk inhoudelijk gerelateerd zijn. Deze procedure noemen we Latente Klasse Bomen (LKB) analyse.

LKB modeleren begint als een standaard LK analyse door te bepalen of een model met twee klassen een betere beschrijving van de data geeft dan model met één klas. Als het twee klassen model wordt geprefereerd dan worden de twee klassen apart genomen en voor iedere klas wordt een nieuwe data set wordt geconstrueerd. De respondenten worden proportioneel toegewezen aan iedere klas. Vervolgens wordt voor ieder van deze datasets opnieuw getest of één klas of twee klassen geprefereerd worden. Hierdoor ontstaat een hiërarchische boomstructuur van latente klassen. Het grootste voordeel van deze aanpak is dat klassen dan gemakkelijk inhoudelijk aan elkaar te relateren zijn. Hierdoor kan bij een LKB zowel statistische als inhoudelijke argumentatie gebruikt kan worden om het aantal klassen te bepalen. Deze LKBs met twee splitsingen is beschreven in hoofdstuk twee en de praktische toepassing is geïllustreerd door middel van een empirisch voorbeeld over sociaal kapitaal.

De eerste splitsing van een LKB is het belangrijkst omdat de rest van de boom hiervan afhankelijk is. Bovendien zijn er situaties waarin een binaire splitsing niet voldoet en het is belangrijk om dit te kunnen bepalen. Daarom is in hoofdstuk 3 een maat voor relatieve verbetering van fit voorgesteld om modellen met een verschillend aantal klassen te vergelijken. Standaard fit maten zouden resulteren in standaard LK analyse, maar met de relatieve fit verbetering kan worden onderzocht of het geprefereerd moet worden om meer dan twee klassen te gebruiken bij de eerste splitsing. Bij volgende splitsingen is inhoudelijke informatie over de klassen aanwezig die kan helpen om het aantal klassen te bepalen, terwijl de relatieve fit verbetering ook opnieuw toegepast kan worden. Met empirische voorbeelden over sociaal captiaal en (post-)materialisme zijn verschillende LKBs geïllustreerd die gebaseerd zijn op de relatieve fit maat.

LK analyse wordt in principe toegepast op cross-sectionele data (data die op één moment is gemeten). Voor longitudinale data worden latent klasse groei curve analyse gebruikt. Deze analyse identificeert gelijke patronen over tijd, zoals mensen met vergelijkbaar stemmingsverloop gedurende een dag. Een groot aantal klassen is vrij gebruikelijk bij longitudinale data en de bomen aanpak kan zeer toepasselijk zijn voor het vaststellen van verschillende patronen over tijd. Daarom is in hoofdstuk vier de LKB procedure uitgebreid zodat er ook Latente Klasse Bomen (LKGB) geconstrueerd kunnen worden. Aan de hand van empirische voorbeelden over stemming gedurende de dag en de kans op drugsgebruik gegeven iemands leeftijd is de LKGB procedure geïllustreerd.

Het vaststellen van de klassen is vaak de eerste stap bij het analyseren van onderzoekgegevens. Vervolgens willen onderzoekers graag de klassen relateren aan externe variabelen. Zo wil men bij voorbeeld testen of het aantal mannen en vrouwen verschilt tussen de klassen of dat wellicht klasse lidmaatschap kan worden voorspeld aan de hand van een bepaalde factor zoals leeftijd. In hoofdstuk vijf is een procedure ontwikkeld die de verdeling van een variabele vergelijkt tussen de klassen en die klasse lidmaatschap voorspelt aan de hand van externe variabelen. Dit wordt toegepast op iedere split van een boom en geeft daardoor een duidelijk overzicht van hoe iedere klasse van de boom gerelateerd is aan de externe variabelen. Deze procedure is geïllustreerd voor zowel LKB als LKGB in hoofdstuk vijf.

In het laatste hoofdstuk worden mogelijkheden voor vervolgonderzoek besproken, zoals steekproeven binnen een steekproef trekken en het effect van verschillende fit maten. In principe zal de traditionale LK analyse nog steeds een eerste uitgangspositie van een toegepast onderzoeker zijn, maar met dit proefschrift (waaronder ook het ontwikkelen van software die de bomen bouwt valt) is er de mogelijkheid om in lastige situaties over te stappen naar de LKB methode, zodat inhoudelijke argumentatie gebruikt kan worden om het aantal klassen te bepalen.

Acknowledgements

I am quite proud of this little book, but I did not accomplish this on my own: I had help and support of many people. I mention everyone only once here, though some people belong in several sections. Those people know who they are and I can assure them that I am gratefull to them all for every aspect that helped me get to this point.

First of all I would like express my gratitude to my promotor, Jeroen Vermunt. You made this PhD-project possible. A long time before my project started you made sure that latent class analysis could be thoroughly investigated and I am very happy that I was allowed as the last PhD on your VICI-project. You were always open to help me if I got stuck and after a discussion with you the world often looked very clear and simple (until I had to write it down myself an hour later). I look forward to working together also the upcoming years. Though Jeroen initiated and conceived the project, it was my co-promotor, Verena Schmittmann that matched me with the project. Without you I probably never would have ended up in Tilburg and I would have known a lot less about latent class analysis. Furthermore, I would like to thank my committee for reading, commenting and approving my thesis.

In Tilburg there were two other guys working on latent class analysis. Erwin and Geert, I'm sorry that I could not always laugh at your jokes, but I do think we had some very good times. Many discussions on job related topics definitely expanded my knowledge, but I also thoroughly enjoyed the talks on not job related subjects. We rocked some great conferences together, and discovering New York with you guys was one of the highlights of my PhD. Geert, we really collaborated well, which resulted in Chapter 3 of this thesis. Furthermore, you are a intriguing man and if I ever become half of the loving father you are, then I haven't done very bad. I wish you all the best with chemometrics in Nijmegen. Erwin, you basically always manage to have a refreshing look at whatever subject we discuss and you're really a pleasure to have around. I am very happy that you have joined me on the eighth floor, where you support me in many ways almost every day. Erwin en Geert, thank you for doing me the honor of being my paranimfs.

It is quite a travel from Amsterdam to Tilburg so often, but luckily the department is filled with great people, which eases the burden a lot. Of these many wonderful colleagues, I got a room with *der* Dino. Though we were hardly ever on the same time schedule, we have managed working hard and hardly working very well. My efforts to tell you to come at 9:30 at the latest will probably remain unfruitful (unless you have an early appointment), but I will stay persistent until the end of your PhD. The final year we changed rooms and Jaap Joris joined us. JJ, I bet you're one of the best teachers to have at the entire university. You're a great and very versatile guy and I can´t thank you enough for the times that you have driven me to and from Den Bosch (spam-alert: for a great photographer, check www.superformosa.nl). Dino and JJ, I am very happy with the two of you as my office mates and hope that the three of us stay put in S806 for a long time.

I am also much obliged to everybody that participated in the research groups over the years. In the order that they got their PhD/I think they will get their PhD (we'll see how that works out): kalme Katrijn, dromerige Daniel, keurige Kim, mooie Margot, zsuper Zsuza, daredevil Dereje, lenige Lianne, levendige Leonie and noob Niek, thank you all for helping me improve my papers and also for letting me criticize your work, which has taught me a lot! Besides doing my own research, I also had some teaching obligations. Teaching Qualitative Research Methods started relatively easy with intrigerende Ingrid and charmante Coosje. However, after an evaluation showed that the program had to be intensified, I had to work extremely hard with handige Hilde and lijpe Laura. Nowadays pittige Pia has joined the team and the course is very manageable. Ladies, thank you for learning and teaching about qualitative methods with me, and stralende John, thank you for teaching us. Some colleagues also lightened the load of traveling up and down to Tilburg, by travelling with me. The NS does not always cooperate, but delays and detours were less troublesome if I ran into representatieve Robert, jeugdige Jesper, remarkable Reza, rocking Rolien, precieze Pieter, jennende Joris or (fouten)jager Jelte. I have mentioned a lot of people from Tilburg already, but many also supported me by making me feel at home in Tilburg. Weergaloze Wilco, malle Marcel, imposante Inga, florissante Florian ("auf Deutschland!"), dansende Davide (thanks for lending me your couch several times!), prachtige Paulette, sportambassadeur Robbie, lieftallige Eva, charismatische Chris, smoking Sara, eclatante Elise en jolige Jules, thanks for every time you guys climb all those stairs to the 8th floor to enjoy your lunch. I also want to mention the nestors lollige Luc and genereuze Guy, who have been less present lately, which saddens me. I hope to see the two of you back in the office when you're completely recovered. Finally, I would like to thank machtige Marieke and attente Anne-Marie without whom the entire department would probably fall apart.

Outside of Tilburg I also had quite some support. First, I would like to give a shout-out to the psychological methods department in Amsterdam. People like Han van der Maas, Denny Borsboom, Conor Dolan, Dylan Molenaar, Marijke Engels-Freeke and Harry Vorst have taught me a lot and provided an awesome educational program. But foremost, the time I spend there with my fellow students convinced me to take the academic path I am taking now. Sacha, Alexander, Vera, Michèle, Lotte, Marie, Sjoerd, Cesar, Anja, Paul, Esther, Suzan, Ravi and Daan, we had an amazing time becoming real methodologists. The greatest loss during my PhD has been the passing of Janneke, who was not only one of the most promising scientists I have ever met, but also a sweet, crazy, creative and dear friend.

Outside of academia I also had a lot of mental and social support. Cees, I admire your sincerity and you seem to be always genuinely interested, also in boring statistics talks. Bas, Yarin, XM, Philo, Yodit, Joris, and the other Ignatianen, high school is a long time ago, but we try to see each other at least a few times a year. Friendships with a strong foundation as ours is priceless. My teammates of het Roze Legioen also supported me by learning how to handle dissapointments when we lost yet another match. Furthermore, my study friends who did not share my passion for methodology, but are nevertheless very nice people: Philip, Luuk, Tiel, Myrthe en Maaike. The last years we have seen each other less than during our studies, but any time we manage to meet up, it feels like nothing has changed. I also want to mention Marloes and Casper with whom I had some very good company and I especially would like to thank Casper for designing the cover of this book.

Mijn laatste dankwoorden zijn voor mijn familie. Sinds een aantal jaren is deze twee keer zo groot, doordat Esther, Gotam, Stefan, Arianne, Michael, Truus, Doritha en Erwin mij in hun midden hebben opgenomen: *Vielen Dank und ich hab euch alle lieb!* Mijn lieve ouders, dankzij jullie heb ik een geweldig onbezorgd leven, jullie onvoorwaardelijke steun is waanzinnig. Pap, ik weet dat je graag grappend zegt dat je beide zonen hun grote voorbeeld hebben gevolgd en hoewel Don en ik dat nooit hardop zullen toegeven, heb je toch wel verdomd veel bewijs voor deze hypothese. Mam, je hebt het niet altijd gemakkelijk met mij (en vooral in combinatie met pappa en Don), maar jij laat je niet snel uit het veld slaan. Je bent één van de sterkste mensen die er zijn (ondanks je lengte), de allerliefste moeder en ik bewonder je absoluut net zo veel als die persoon die denkt mijn voorbeeld te zijn. Don, ik vind het ontzettend leuk hoeveel interesses wij delen, maar je hebt me toch een aantal keren behoorlijk verbaasd. Ik zag nog wel aankomen dat jij beter zou worden met R dan ik, maar ik had nooit verwacht dat zelfs onze geliefden uit dezelfde omgeving zouden komen. Een goede smaak heb je in ieder geval wel, want ik ben heel blij met Selina als mijn Schwägerin. Helen, ik weet dat je ontzettend trots bent je op je zogenaamd slimme broertje, maar dit broertje is ook ontzettend trots op zijn waanzinnig intelligent zus die ook nog eens heel lief is. Peter, jij en Helen hebben het ontzettend leuk, het is altijd gezellig met jou en ik ben je dankbaar voor alle keren dat ik van jouw advies en enorme handigheid gebruik heb mogen maken.

En dan ten slotte, lieve Dorina, ik wil jou graag bedanken voor het feit dat met jou in de buurt altijd de vogeltjes fluiten en vrijwel elke tegenslag mij niet kan deren. Ik denk dat ik dit gevoel niet goed in woorden kan uitdrukken, maar zal proberen het duidelijk te maken terwijl we samen oud worden.