Hierarchical Latent Class Modeling as a Density Estimation Method for Categorical Data

Daniël W. van der Palm, L. Andries van der Ark, and Jeroen K. Vermunt

Correspondence Address:

Daniël W. Van der Palm

Department of Methodology and Statistics FSW

Tilburg University

PO Box 90153

5000 LE Tilburg

The Netherlands

Tel: +31-4663270

Fax: +31-4663002

E-mail: d.w.vdrpalm@uvt.nl

Abstract

Traditionally latent class (LC) analysis is used by applied researchers as a tool for identifying substantively meaningful clusters. More recently, LC models have also been used as a density estimation tool for categorical variables. We introduce a hierarchical LC (HLC) model as a density estimation tool that may offer several advantages in comparison to a standard LC model. When using an LC model for density estimation, a common model-fit strategy is to estimate increasingly large LC models in terms of the number of latent classes. In practice this strategy may be very time consuming or even unfeasible, especially if a dataset with a large number of variables is analyzed. A considerable number of increasingly large LC models may have to be estimated before sufficient model-fit is achieved. An HLC model consists of a hierarchical sequence of small LC models. Therefore, an HLC model can be estimated much faster and is less demanding in terms of computer working-memory, meaning that this model is more widely applicable and practical. In this study we describe the algorithm of fitting an HLC model, and discuss the various settings that indirectly influence the precision of an HLC model as a density estimation tool. These settings are examined in a simulation study, and the best performing algorithm is applied to a real-data example in the context of missing data and multiple imputation. Results from a generated data example show that an HLC model is able to correctly model complex association among categorical variables.

Keywords: Latent class analysis, categorical data, mixture model, density estimation, hierarchical latent class model, missing data.

1. Introduction

Traditionally, latent class (LC) analysis (Lazarsfeld, 1950; also see, e.g., McCutcheon, 1987; Goodman, 1974; Hagenaars and McCutcheon, 2002; Collins and Lanza, 2010) is used as a statistical method to identify substantively meaningful groups from multivariate categorical data (e.g., Muthén, 2004; Magidson and Vermunt, 2004). Some examples include the empirical definition of eating disorder subgroups (Keel et al., 2004), empirical identification of discrete subgroups with similar symptoms of posttraumatic stress disorder (Breslau, Reboussin, Anthony, and Storr, 2005), exploring whether certain subtypes of Antisocial Personality Disorder exist (Bucholz, Hesselbrock, Heath, Kramer, and Schuckit, 2000), and unsupervised learning of the meaning of words (Hofmann, 2001). Typically, a substantive interpretation is given to the LCs based on the estimated model parameters. To facilitate interpretation, it is desirable that the number of LCs is small, the LC model is identifiable (e.g., Goodman, 1974), and the global maximum of the likelihood has been found.

More recently, LC models have been used in a different way: as estimators of the joint density of a set of categorical variables. The often complex multivariate density is approximated by a finite mixture of simpler multinomial densities; examples can be found in various fields. Vermunt, Van Ginkel, Van der Ark, and Sijtsma (2008; also, see Gebregziabher and DeSantis, 2010, and Van der Palm, Van der Ark, and Vermunt, 2012) used the estimated density for multiple imputation of incomplete categorical data. In the context of the analysis of voting behavior, Linzer (2011) proposed using an LC model for smoothing large contingency tables with many zero observed frequencies. He showed that analyzing the table with the estimated frequencies from an LC model instead of the observed frequencies may greatly improve the quality of the obtained results. In the context of psychological testing, Van der Ark, Van der Palm, and Sijtsma (2011) estimated the density of item-score vectors obtained from psychological test data using LC analysis, and derived a coefficient for test-score reliability. In the context of pattern recognition, Bouguila and ElGuebaly (2009) showed that an LC model can be used to summarize an image database by estimating the density of image characteristics.

The idea of approximating a complex density by a mixture of simpler densities is well-known in finite mixture modeling (e.g., McLachlan and Peel, 2000, pp. 11-14) but the majority of research

has focused on mixtures of continuous distributions (e.g., Everitt, Landau, and Leese, 2001, pp. 8-10). The most important issue when using LC models to estimate densities is the precision of the density estimate. Depending on the application of interest, the two-way, three-way, or higher-way interactions should be accurately described by the LC model. In this context, the LC model is solely used as a tool, and the interpretation of the latent classes is not important. Consequently, issues that play an important role in interpreting LCs—number of LCs and identifiability—are hardly important when LC models are used to estimate densities. Vermunt et al. (2008) argued that over-fitting the data has no impact on the precision of the density estimate. Therefore, the number of LCs is of less importance, as long as the model is able to yield a precise density estimate.

For datasets containing a large number of variables, a large number of LCs is usually required for precise density estimation. Let LC(k) denote an LC model with k classes. For example, Vermunt et al., (2008) used AIC (Akaike, 1974) as a criterion and selected an LC(50) model to model a survey dataset of 79 variables. They indicated that even more LCs may have been needed for precise density estimation. A typical model-fit strategy is to estimate an LC(5), LC(10), LC(15), LC(20) model, etcetera, until the model fit no longer improves. This can be a very time-consuming process: For example, we reanalyzed the survey data set used by Vermunt et al., containing 4292 cases and 79 categorical variables (for details, see Mittelhaëuser, Van der Ark, and Richards, 2010), and estimated an LC(5), LC(10), LC(15),…, LC(60), and LC(65) model. The analysis took approximately 8 hours and 12 minutes (details in Table 1) on a, for current standards, very fast personal computer (i7 2600 quadcore processor, 8GB of internal memory), as summarized in Table 1. As overfitting is not problematic, an LC(65) model may be taken as the final solution, as it yields the first considerable increase in AIC (Table 1). The long computation time and comparison of many LC models can be an obstacle for researchers, especially when a density has to be estimated multiple times (e.g. multiple imputation based on bootstrap replications). Hence, for larger data sets, density estimation using traditional LC models is problematic.

- Insert Table 1 about here -

In this paper we introduce the hierarchical LC (HLC) model as a fast alternative to the LC model for density estimation. The HLC model requires only a single run. An HLC model is obtained by fitting a

sequence of LC(1) and LC(2) models. A key characteristic of this sequence of models is that is has a hierarchical structure; every analysis builds on the results of the previous analyses. An exemplary HLC model is shown in Figure 1 to illustrate its structure.

- Insert Figure 1 about here -

At the top, the hierarchy starts with the full sample and, at the bottom, the hierarchy ends in a number of LCs, each containing a portion of the sample. We refer to these LCs as endpoints, and depict them as LCs printed in bold. In this example, a total of 11 classes were considered for splitting (numbered as nodes 1 through 11); in four instances an LC(2) model sufficiently improved model fit compared to an LC(1) model (classes at nodes 1, 2, 3, 5, and 8), resulting in a split. At the other nodes, an LC(2) model did not sufficiently improve model-fit, which yields the 6 endpoints (classes 4, 6, 7, 9, 10, and 11). As can be seen, every split is a step down in a hierarchy consisting of 5 levels for this example. The notation used to denote the LCs at the different levels is explained in the HLC section.

A key characteristic of a standard LC model is that each analysis starts from scratch. That is, when fitting a model with $K + 1$ classes, the information obtained with the K-class model is entirely neglected. The HLC model is able to reduce computation time because it is a stepwise procedure and each step takes into account the information obtained in the previous steps; the density estimate yielded by an HLC model consists of a sequence of hierarchically linked LC(2) models. Furthermore, in contrast to the standard LC model, the number of LCs is not specified a priori for an HLC model; the number of LCs is increased during the estimation process until a sufficiently precise density estimate is obtained. Therefore, it is also no longer necessary to estimate and compare several models.

Due to its hierarchical nature and relative simplicity we expect the estimation of an HLC model to be faster than of a traditional LC model, controlling for computer hardware. In addition, it is much easier to utilize multiple processor cores in the estimation of an HLC model (to be explained in the HLC section). The remainder of this paper is organized as follows. First, we provide a full description of the HLC model and explain why estimating the HLC model is fast. Second, we discuss different choices that can be made in the construction of an estimation algorithm. Third, using a

generated data example, we compare the effect of these different choices on the precision of estimating complex densities. Fourth, the best performing estimation algorithm is applied to a dataset that was also analyzed by Vermunt et al. (2008) using a standard LC model and we compare the results.

2. Hierarchical Latent Class Model

An HLC model, which could also be referred to as a hierarchical mixture model, is fitted to the data in a stepwise manner. From the starting point towards the endpoints, a branching pattern is observed (Figure 1), which is established as follows: first (level 0), the whole sample is analyzed to assess whether an LC(2) model provides a better fit than an LC(1) model, according to specific decision rules (to be discussed shortly). If this is the case, two LCs are created, effectively splitting the whole sample into two subsamples (level 1). Next, the two subsamples in level 1 are analyzed separately, to assess whether an additional split further improves the overall model-fit. This procedure continues until a split does not improve the model-fit sufficiently; an endpoint has been reached in the hierarchy. During the analysis of each class, the proportion of the sample that was already assigned to other LCs is held constant; this is sometimes referred to as partial EM (Ueda and Nakano, 2000; Wang, Luo, Zhang, and Wei, 2004).

We will now give a technical description of the HLC model and the steps that need to be taken to obtain a density estimate. Let $x_{r|\mathbf{x}_{r-1}} \in \{1,2\}$ denote a binary latent variable at level *r* (see Figure 1) in the hierarchy. Each outcome of $x_{r|\mathbf{x}_{r-1}}$ (i.e., $x_{r|\mathbf{x}_{r-1}} = 1$ and $x_{r|\mathbf{x}_{r-1}} = 2$) is an LC. Vector $\mathbf{x}_{r-1}$ is called the lineage vector and gives the sequence of 'parent' LCs at levels $1,\ldots,r-1$. In general,

$$\begin{aligned}\mathbf{x}_{r-1} &= \left(x_1, x_{2|x_1}, x_{3|x_1,x_2}, \ldots, x_{r-1|x_1,x_2,\ldots x_{r-2}}\right) \\ &= \left(x_1, x_{2|x_1}, x_{3|\mathbf{x}_2}, \ldots, x_{r-1|\mathbf{x}_{r-2}}\right)\end{aligned}.$$

For example, let $x_1 = 1$ be the parent LC for the binary latent variable $x_{2|1}$ at the second level; hence, *r*=2, and the lineage vector $\mathbf{x}_{r-1} = 1$. The two LCs of this variable are denoted $x_{2|1} = 1$ and $x_{2|1} = 2$, respectively. Then, let $x_{2|1} = 1$ be the parent LC for the binary latent variable $x_{3|1\,1}$ at the third level;

hence, *r*=3, and the lineage vector $\mathbf{x}_{r-1} = (1,1)$. The two LCs of this variable are denoted $x_{3|1\,1} = 1$ and $x_{3|1\,1} = 2$. This procedure continues until splitting an LC no longer improves the overall model-fit. Figure 1 shows the LC names for the example HLC model.

To be able to discuss all the possible steps of fitting an HLC model we must define a general notation for an LC in the hierarchy at an arbitrary level and with an arbitrary lineage of parent LCs. Such a class is simply denoted as $x_{r|\mathbf{x}_{r-1}}$, hence, the level and lineage vector are unspecified. To illustrate the use of this notation we give an example: if we split LC $x_{r|\mathbf{x}_{r-1}} = 1$, LCs $x_{r+1|\mathbf{x}_r} = 1$ and $x_{r+1|\mathbf{x}_r} = 2$ are created, both with an unspecified lineage vector $\mathbf{x}_r$. Figure 2 depicts the three possible situations that can occur when fitting an HLC model: the first split of the data, a second split, and the *r*th split.

- Insert Figure 2 about here -

2.1 The first split

Let $\mathbf{y}_i = \left(y_{i1}, ..., y_{ij}, \ldots, y_{iJ} \right)$ denote the vector of scores on the J variables for respondent *i*. The first split consists of fitting an LC(1) and LC(2) model on the total sample of size *n*. The LC(1) model is simply the independence model, with log-likelihood,

$$\log L_1 = \sum_{i=1}^{n} \log \prod_{j=1}^{J} P(y_{ij}) . \qquad (1)$$

Let $P(x_1), x_1 \in \{1,2\}$, denote the LC proportions, and let $P(y_{ij} \mid x_1), x_1 \in \{1,2\}$, denote the conditional response probability for variable *j*. The LC proportions and the conditional response probabilities are the parameters of the LC model. In the LC(2) model, the probability of observing score vector $\mathbf{y}_i$ is defined as

$$P(\mathbf{y}_i) = \sum_{x_1=1}^{2} P(x_1) \prod_{j=1}^{J} P(y_{ij} \mid x_1) . \qquad (2)$$

Responses within each LC are assumed to be independent; this is generally referred to as the local independence assumption (Lazarsfeld, 1950). Equation 2 yields the following log-likelihood for the LC(2) model:

$$\log L_2 = \sum_{i=1}^{n} \log \sum_{x_1=1}^{2} P(x_1) \prod_{j=1}^{J} P(y_{ij} \mid x_1) . \qquad (3)$$

We obtain maximum-likelihood estimates of $P(x_1)$ and $P(y_{ij} \mid x_1)$ in the same way as for standard LC analysis, using an EM algorithm (Dempster, Laird, and Rubin, 1977), a Newton-Raphson procedure, or a combination of the two. If $\log L_2$ is sufficiently larger than $\log L_1$, the total sample is split.

2.2 A second split

After the first split, all respondents are divided between LCs $x_1 = 1$ and $x_1 = 2$, effectively creating two subsamples. For each of these two subsamples we estimate an LC(1) and LC(2) model, and decide whether these LCs should be split up again. However, to analyze a subsample, it must be known which respondents belong to it. To this end, we use the posterior membership probabilities $P(x_1 \mid \mathbf{y}_i)$, which are defined as follows,

$$P(x_1 \mid \mathbf{y}_i) = \frac{P(x_1, \mathbf{y}_i)}{P(\mathbf{y}_i)} = \frac{P(x_1) \prod_{j=1}^{J} P(y_{ij} \mid x_1)}{\sum_{x_1=1}^{2} P(x_1) \prod_{j=1}^{J} P(y_{ij} \mid x_1)} . \qquad (4)$$

Full LC membership of one of the two LCs could be forced for each person by modal assignment or a random assignment using the posterior membership probabilities defined in Equation 4. However, to stay as close to the estimated model as possible, and to reflect the uncertainty about LC membership, each respondent is included in both subsamples with a weight equal to the posterior membership probability. Hence, respondent *i* receives weight $P(x_1 = 1 \mid \mathbf{y}_i)$ for the analysis of the first LC and weight $P(x_1 = 2 \mid \mathbf{y}_i) = 1 - P(x_1 = 1 \mid \mathbf{y}_i)$ for the analysis of the second LC.

We now consider LC $x_1 = 1$ (Figure 2, 2nd split). If splitting this LC sufficiently improves model fit, latent variable $x_{2|1}$ is created with LCs $x_{2|1} = 1$ and $x_{2|1} = 2$. To decide whether LC $x_1 = 1$

should be split, we estimate an LC(1) and LC(2) model for the subsample in $x_1 = 1$. The contribution of each person to the likelihood of the LC(1) and LC(2) model must be weighted by the probability of belonging to the subsample under consideration. Therefore, a weighted log-likelihood equation is defined, with the posterior membership probabilities as weights. For the LC(1) model it is defined as

$$\log L_1(x_1 = 1) = \sum_{i=1}^{n} P(x_1 = 1 \mid \mathbf{y}_i) \log \prod_{j=1}^{J} P(y_{ij}), \tag{5}$$

and for the LC(2) model it is defined as

$$\log L_2(x_1 = 1) = \sum_{i=1}^{n} P(x_1 = 1 \mid \mathbf{y}_i) \log \sum_{x_{2|1}=1}^{2} P(x_{2|1}) \prod_{j=1}^{J} P(y_{ij} \mid x_{2|1}). \tag{6}$$

Once again, the maximum (weighted) likelihood estimates of $P(x_{2|1})$ and $P(y_{ij} \mid x_{2|1})$ are obtained, and if $\log L_2(x_1 = 1)$ is sufficiently larger than $\log L_1(x_1 = 1)$, LC $x_1 = 1$ is split.

2.3 The r*th split*

We now describe the most general situation, the *r*th split, and consider $x_{r|\mathbf{x}_{r-1}} = 1$ (See Figure 2). To do so, the posterior class membership probabilities $P(x_{r|\mathbf{x}_{r-1}} = 1 \mid \mathbf{y}_i)$ must be calculated as these are the weights defining the subsample at this node. It should be noted that the LC(2) model estimated at the previous node yielded $P(x_{r|\mathbf{x}_{r-1}} = 1 \mid \mathbf{y}_i, x_{r-1|\mathbf{x}_{r-2}})$, the posterior probability of belonging to LC $x_{r|\mathbf{x}_{r-1}} = 1$ given $\mathbf{y}_i$ and its 'parent' LC $x_{r-1|\mathbf{x}_{r-2}}$; we refer to this probability as a local posterior membership probability, as it only holds for the members of $x_{r-1|\mathbf{x}_{r-2}}$. Figure 3 is an elaboration of Figure 2, and shows for each LC how the posterior membership probabilities can be obtained from the local posterior membership probabilities.

- Insert Figure 3 about here -

Note that for level 1, the first level in which weights are used, the local posterior membership probabilities and posterior membership probabilities are identical. For the next levels, the posterior

membership probabilities are computed by taking the product of all the local posterior membership probabilities that are associated with the lineage of parent LCs. Hence,

$$P(x_{r|\mathbf{x}_{r-1}}=1|\mathbf{y}_i)=P(x_1|\mathbf{y}_i)\prod_{q=2}^{r}P(x_{q|\mathbf{x}_{q-1}}|\mathbf{y}_i,x_{q-1|\mathbf{x}_{q-2}}).$$

Once the posterior membership probabilities are defined, the calculations for the *r*th split are very similar to those performed for a second split. The weighted log-likelihood equation of the LC(1) model for the *r*th split is defined as

$$\log L_1(x_{r|\mathbf{x}_{r-1}}=1)=\sum_{i=1}^{n}P(x_{r|\mathbf{x}_{r-1}}=1|\mathbf{y}_i)\log\prod_{j=1}^{J}P(y_{ij}) \tag{7}$$

and the weighted log-likelihood of the LC(2) model as

$$\log L_2(x_{r|\mathbf{x}_{r-1}}=1)=\sum_{i=1}^{n}P(x_{r|\mathbf{x}_{r-1}}=1|\mathbf{y}_i)\sum_{x_{r+1|\mathbf{x}_r}=1}^{2}P(x_{r+1|\mathbf{x}_r})\prod_{j=1}^{J}P(\mathbf{y}_{ij}|x_{r+1|\mathbf{x}_r}). \tag{8}$$

If $\log L_2(x_{r|\mathbf{x}_{r-1}}=1)$ is sufficiently larger than $\log L_1(x_{r|\mathbf{x}_{r-1}}=1)$, LC $x_{r|\mathbf{x}_{r-1}}=1$ is split.

To investigate the next (r+1[th]) split, the estimates of $P(x_{r+1|\mathbf{x}_r})$ and $P(y_{ij}|x_{r+1|\mathbf{x}_r})$ are used to calculate the local posterior membership probabilities for the two newly created LCs, $x_{r+1|\mathbf{x}_r}=1$ and $x_{r+1|\mathbf{x}_r}=2$, as follows:

$$P(x_{r+1|\mathbf{x}_r}|\mathbf{y}_i,x_{r|\mathbf{x}_{r-1}}=1)=\frac{P(x_{r+1|\mathbf{x}_r})\prod_{j=1}^{J}P(y_{ij}|x_{r+1|\mathbf{x}_r})}{\sum_{x_{r+1|\mathbf{x}_r}=1}^{2}P(x_{r+1|\mathbf{x}_r})\prod_{j=1}^{J}P(y_{ij}|x_{r+1|\mathbf{x}_r})}. \tag{9}$$

The posterior membership probabilities defining the partial membership weights of the two new LCs, $x_{r+1|\mathbf{x}_r}=1$ and $x_{r+1|\mathbf{x}_r}=2$, are obtained as

$$P(x_{r+1|\mathbf{x}_r}|\mathbf{y}_i)=P(x_{r|\mathbf{x}_{r-1}}=1|\mathbf{y}_i)P(x_{r+1|\mathbf{x}_r}|\mathbf{y}_i,x_{r|\mathbf{x}_{r-1}}=1).$$

2.4 Decision Rules

At every step in the hierarchy it must be decided whether the targeted class should be split further; that is whether splitting the class concerned sufficiently improves model-fit. For the initial split we base this decision on the difference between the log-likelihood functions defined in Equations 1 and 3. For a second and the rth split we compare the weighted log-likelihood function of Equation 5 with Equation 6 and of Equation 7 with Equation 8, respectively. However, the question remains what constitutes a sufficient increase in log-likelihood. This is the most important setting of an HLC model as it is expected to greatly affect the precision of the density estimate.

A commonly used approach to assess model-fit for LC models is to use a relative fit statistic such as the Bayesian information criterion (BIC; Schwarz, 1978), AIC, or a variant of AIC called AIC3 (Bozdogan, 1993). These relative fit statistics combine model fit (log-likelihood, LL) and parsimony (number of parameters). Several simulation studies have shown that BIC tends to select too few LCs and AIC tends to selects too many LCs (Lin and Dayton, 1998; Andrews and Currim, 2003); AIC3 was found to be a good compromise between BIC and AIC in terms of the selected number of LCs (Andrews and Currim, 2003; Dias, 2006). However, because model under-fit can lead to serious bias and model over-fit does not, AIC is preferred over BIC and AIC3 in the context of density estimation (e.g. Vermunt et al., 2008). Moreover, it could be argued that parsimony should not be taken into account at all when using an LC model as a density estimator. Therefore, we also consider a decision rule that is only based on the difference in the log-likelihoods of an LC(1) model and LC(2) model.

If the size of a class is too small in relation to the number of parameters of the LC models, it is not known whether the density estimate will still improve if we continue splitting LCs, or whether the density estimate may even deteriorate. Therefore, we also included several rules with regard to minimal class size.

A program was written in C++ which performs all the necessary steps in the estimation of a HLC model, and it is available on request. For the estimation of the LC(1) and LC(2) models at a specific split, it runs LatentGold 4.5 (Vermunt and Magidson, 2008) in batch mode using an appropriately weighted data set.

3. Generated data example

To investigate the precision of an HLC model as a density estimator, we applied the model at population level. By doing so, we can focus on the precision of an HLC model without considering the influence of sampling error. We defined a complex population model (depicted in Figure 4) for 11 dichotomous variables and obtained the population probabilities for each response pattern. By multiplying these population proportions by 1000 we obtained frequencies for all the response patterns, amounting to a sample (of size N=1000) that is exactly in accordance with the population. The precision of an HLC model is defined as the difference between the predicted frequencies for each response pattern and the population frequencies.

- insert Figure 4 about here -

The population model is a path-model for categorical data (Goodman, 1973) consisting of two sets of independent variables (y_1, y_2, y_3 and y_4, y_5, y_6, y_7, y_8), and 3 dependent variables y_9, y_{10}, and y_{11}. By using this population model it can be assessed whether an HLC model is able to estimate the density of the data from a complex population (i.e. containing three-way associations) and having no LC structure. Let β_j denote a log-linear parameter value for the *j*th variable. The density of y_1, y_2, and y_3 is defined as,

$$\log P(y_{i1}, y_{i2}, y_{i3}) = \sum_{j=1}^{3} \beta_j^1 y_{ij} + \sum_{j=1}^{2} \sum_{j'=j+1}^{3} \beta_{jj'}^2 y_{ij} y_{ij'} + \beta_{123}^3 y_{i1} y_{i2} y_{i3} .$$

Hence, the joint density of y_1, y_2, and y_3 is in agreement with a saturated log-linear model containing all one- two- and three-variable associations. The density of y_4, y_5, y_6, y_7 and y_8 is defined as

$$\log P(y_{i4}, y_{i5}, y_{i6}, y_{i7}, y_{i8}) = \sum_{j=4}^{8} \beta_j^4 y_{ij} + \sum_{j=4}^{7} \sum_{j'=j+1}^{8} \beta_{jj'}^5 y_{ij} y_{ij'} ,$$

and only contains one- and two-way associations. The log-linear parameter values for the two models are given in Appendix A.

The conditional probabilities of the three dependent variables are defined to be in agreement with logit models, using effects coding for the parameters. Let β_j^q denote a logit regression parameter for the regression of dependent variable *q* on the *j*th independent variable. For dependent variable y_9,

$$\text{logit}(y_9) = \beta_0^{y9} + \beta_1^{y9} y_1 + \beta_2^{y9} y_2 + \beta_3^{y9} y_3 + \beta_{12}^{y9} y_1 y_2 + \beta_{13}^{y9} y_1 y_3 + \beta_{23}^{y9} y_2 y_3 + \beta_{123}^{y9} y_1 y_2 y_3,$$

for dependent variable y_{10},

$$\text{logit}(y_{10}) = \beta_0^{y10} + \beta_4^{y10} y_4 + \beta_5^{y10} y_5 + \beta_6^{y10} y_6 + \beta_7^{y10} y_7 + \beta_{67}^{y10} y_6 y_7 + \beta_9^{y10} y_9,$$

and for dependent variable y_{11},

$$\text{logit}(y_{11}) = \beta_0^{y11} + \beta_7^{y11} y_7 + \beta_8^{y11} y_8 + \beta_{10}^{y11} y_{10}.$$

These relationships yield a complex density including three-way associations. Values of the logistic regression parameters for the three dependent variables are given in Appendix B.

We compared the true proportions and the estimated proportions by the HLC model for three combinations of variables: $(y_9 y_{10})$, $(y_8 y_{11})$, and $(y_6 y_7 y_{10})$. Variables y_6, y_7 and y_{10} have a three-way association, and it is important to determine whether an HLC model is able to correctly pick up this complex association. The estimated proportions can easily be calculated from the estimated HLC parameters. For example, the estimated proportions of variables y_6, y_7, and y_{10}, can be obtained as follows,

$$\hat{P}(y_6 y_7 y_{10}) = \sum_x \hat{P}(x)\hat{P}(y_6 \mid x)\hat{P}(y_7 \mid x)\hat{P}(y_{10} \mid x).$$

3.1 Implementation of decision rules

We considered the following implementations of the decision rules discussed in the method section: a minimum of 1 point increase in the LL (HLC-1), a minimum of a 5 point increase in the LL (HLC-5), and a decrease of AIC (HLC-AIC), which amounts to a minimum improvement equal to the number of additional parameters. We included the following rules with regard to minimal class size: (1) no restriction, (2) at least 3% of the total sample, and (3) at least 6% of the total sample.

3.2 Precision of density estimation

The precision of density estimation is defined as the difference between the true proportions and the estimated proportions. To quantify this difference we computed Pearson's chi-square statistic:

$$\chi^2 = 1000 \cdot \sum \frac{(\hat{p} - p)^2}{p}.$$

Hence, the chi-square statistic indicates how well an HLC model approximates the true density of the data. We also included the chi-square statistic for the independence model as a reference point.

4. Results

Table 2 shows that a minimal improvement of 1 point in the log-likelihood leads to the most precise results, for all three marginal tables. HLC-5 and HLC-AIC yielded relatively large chi-square statistics for the marginal tables $y_8 y_{11}$ and $y_6 y_7 y_{10}$, regardless of the minimum class size. A minimum class size of 30 only has a very small effect on the precision for the HLC-1 model, but a minimum class size of 60 did deteriorate the results. For HLC-5 only the minimum class size of 60 affected the results, but for HLC-AIC, minimum class size did not have any effect. The minimum class sizes of 30 for HLC-5 and 30 or 60 for HLC-AIC did not affect the results because the minimal required improvement in the log-likelihood was always encountered before the classes became too small.

- Insert Table 2 about here -

From the last row of Table 2 it can be seen that the chi-square statistic for the entire data produced by the HLC-1 (and no minimum class size) is only 0.063% of the chi-square distance for the independence model, in other words, the imprecision in the density estimate is reduced by 99.937%. HLC-5 also seems to perform rather well with a reduction in chi-square of 99.78% compared to the independence model.

5. Real-data Example

In this example we demonstrate one of the possible applications of an HLC model as a density estimation tool. A frequently encountered problem for applied researchers is that data are incomplete. A commonly used strategy to deal with this problem is to analyze only those cases that are complete. However, such a complete case analysis may lead to biased statistical results (Little and Rubin, 2002) and reduced power (Little and Rubin, 2002; Schafer, 1997). An advanced alternative method to deal

with incomplete data is multiple imputation (Rubin, 1987). Multiple imputation consists of creating *m* completed data sets by replacing the missing values in the data with plausible values *m* times. For more information on multiple imputation and the comparison to complete-case analysis, see e.g. Schafer and Graham (2002). Vermunt et al. (2008) showed by means of a simulation study that multiple imputation using an LC model yields unbiased parameter estimates and standard errors. In addition, the method can handle a large number of variables, but, as mentioned before it can be very time consuming for large data sets.

We investigated the performance of multiple imputation based on an HLC model, a standard LC model and complete-case analysis, in terms of model-fit and computation time. We now return to the analysis of the ATLAS data (Mittelhaëuser, Van der Ark, and Richards, 2010), as discussed in the introduction. In table Table 1 we reported computation time and model-fit of various LC models for this large dataset. The ATLAS study addressed topics such as motivations, activities, and impressions of visitors of cultural sites and events. The dataset consisted of 4292 observations and 79 categorical variables: 52 with 2 categories, 1 with 3, 19 with 5, 2 with 6, and 1 with 7, 8, 9, 10 and 17 categories, respectively. Complete information is only available for 794 respondents.

For proper imputation, the incomplete data must be imputed multiple times to account for parameter uncertainty that is due to the missing data (Rubin, 1987); we chose to impute the data 10 times. For every individual dataset a different set of parameter values of the HLC model should be used to reflect parameter uncertainty. We obtained 10 nonparametric bootstrap samples (for details see, eg. Efron and Tibshirani, 1993) and estimated the HLC model (1 point minimal improvement combined with minimal class size of 1%) 10 times, yielding 10 different sets of parameters. These sets of parameters were then used to impute the original data 10 times.

After performing multiple imputation we selected 6 variables for a substantive analysis. One important question in the dataset concerning respondents' motivations for visiting cultural attractions is "I want to find out more about the local culture", answered on a five-point scale ranging from 1 (totally disagree) to 5 (totally agree). We used this variable as the dependent variable in an (adjacent-category) ordinal regression model (Agresti, 2002, pp. 286-288). Table 3 provides detailed

information on the variables used included in the substantive analysis such as the number of cases with a missing value.

- Insert Table 3 about here -

In order to illustrate the effect of using complete-case analysis on the parameter estimates, we estimated two regression models that are identical except the second model includes one additional variable, "Admissing Expenditure" that reduces the number of complete cases from 3950 to 1424. Tables 4 and 5 present the coefficients of the two ordinal regression models, estimated using complete-case analysis, multiple imputation using a standard LC model, and multiple imputation using an HLC model. For the first analysis (Table 4) it can be seen that the parameter estimates using the three models are rather similar, except for some differences in the parameter estimates for education. Because there is only a small proportion of missing values in this analysis, it not surprising that complete-case analysis and the HLC model gave similar results. It is reassuring that for most regression coefficients, the LC and HLC approach gave similar estimates. There are however some small differences in the parameter estimates for education.

- Insert Table 4 about here -

For the second analysis (Table 5) much greater differences in the parameter estimates are observed between complete-case analysis and the LC or HLC approach. The estimates based on an LC and HLC model appear to be relatively stable over the two analyses whereas the results produced with complete-case analysis changed: the estimated coefficients of age, gender, and education nearly doubled and all standard errors became larger. Although we cannot compare the estimates of the three methods to the population values, as we did in the generated data example, the LC and HLC approach appear to perform well in this application. It is reassuring that the results of the HLC and LC approach largely concur with those of complete-case analysis when the proportion of missingness is small and that the estimates are stable across the two analyses. Furthermore, no large differences were found in the results of the two regression models between the LC and HLC approach.

- Insert Table 5 about here -

We end the real-data example with some information on the log-likelihood of the data that was produced by an individual LC and HLC model, and the required computation time (this includes the

required computation time for each method to establish which model fits best). It can be seen from Table 6 that the HLC-1 (min N of 2%) and HLC-1 (min N of 1%) models yield a better fit, in much less time.

- Insert Table 6 about here -

6. Discussion

Within this study we introduced the HLC model as a density estimation tool for categorical data. We described the details of the estimation algorithm, and performed an initial investigation of its statistical properties. More specifically, we have shown in the generated data example that an HLC model is able to pick up two-way and three-way associations from a complex population model. Within our generated data example it was found that a minimum required class size of 0% or 3% yields the most precise density estimate. Furthermore, we should be liberal about the minimum required improvement of the log-likelihood, as the condition in which the smallest improvement was demanded after each split yielded the most precise density estimate.

It is important to note that we have not taken sampling error into account. This means that additional research is required to investigate the relationship between sample size, the minimum improvement in the log-likelihood, and precision of density estimation. However, it has been found in previous research that over-fitting does not pose a big problem when using an LC model for density estimation (e.g. Vermunt, et al., 2008), therefore, the impact of over-fitting is expected to be limited for an HLC model as well. It should also be ascertained whether certain restrictions are required, such as a minimum class size, to prevent over-fitting the data when dealing with a real sample. In addition to over-fitting the data, it is important to investigate the estimated standard errors in relation to sample size, minimal class size, and minimal required improvement in the log-likelihood in an extended simulation study.

The real-data example showed that an HLC model can easily be applied to a dataset with a large number of cases and polytomous variables. For a traditional LC model with 65 LCs it took more than 8 hours to establish the best fitting model for this dataset, whereas an HLC (min N of 1%) model only required 1 hour and 6 minutes. In addition to being faster it yielded a better fit to the data. In a practical sense this makes a substantial difference for researchers that use an LC model as a density

estimation tool. Our exemplary application underlines the benefits of an HLC model. If a researcher wants to use multiple imputation, the density of the data has to be estimated several times (10 times in this case). Hence, using an HLC model for multiple imputation instead of an LC model reduced the runtime for this dataset from 20h12m (8h12m + 10*1h12m) to 11h0m. In this calculation of the computation time for the LC approach, the number of LCs that is used for each bootstrap sample is held constant, whereas for the HLC model the optimal number of LCs is determined for each bootstrap sample. The same approach using an LC model would require more than 82 hours of computation time. Furthermore, as noted in the introduction, it is much easier to make use of multiple processing cores for the estimation of an HLC model than for an LC model. This is due to the fact that every processing core can investigate one possible split; with a quad-core processor, this constitutes running 4 independent partial EM algorithms, whereas for an LC model the processing load would have to be divided within one EM algorithm, which is much more difficult to implement, and less efficient.

Appendix A

Log-linear parameters for the density of y_1, y_2, and y_3.

Parameter	*Value*
β^1_1	*.2*
β^1_2	*-.6*
β^1_3	*.4*
β^2_{12}	*.2*
β^2_{13}	*.6*
β^2_{23}	*-.1*
β^3_{123}	*.2*

Log-linear parameters for the density of y_4, y_5, y_6, y_7 and y_8.

Parameter	*Value*	*Parameter*	*Value*	*Parameter*	*Value*
β^1_4	.2	β^2_{45}	.4	β^2_{57}	.6
β^1_5	-.6	β^2_{46}	-.2	β^2_{58}	-.2
β^1_6	.4	β^2_{47}	.6	β^2_{67}	.1
β^1_7	.2	β^2_{48}	-.3	β^2_{68}	-.2
β^1_8	.6	β^2_{56}	.8	β^2_{78}	.6

Appendix B

Logistic regression parameters for the conditional densities of y_9, y_{10} and y_{11}.

Dependent Variable					
y_9		y_{10}		y_{11}	
Parameter	*Value*	*Parameter*	*Value*	*Parameter*	*Value*
$\beta_0^{y_9}$	.0	$\beta_0^{y_{10}}$	.0	$\beta_0^{y_{11}}$	.0
$\beta_1^{y_9}$	.3	$\beta_9^{y_{10}}$	.5	$\beta_7^{y_{11}}$	-.6
$\beta_2^{y_9}$	.6	$\beta_4^{y_{10}}$	.6	$\beta_8^{y_{11}}$	.2
$\beta_3^{y_9}$	-.9	$\beta_5^{y_{10}}$	.1	$\beta_{10}^{y_{11}}$	.4
$\beta_{12}^{y_9}$	.3	$\beta_6^{y_{10}}$	-.4		
$\beta_{13}^{y_9}$	.5	$\beta_7^{y_{10}}$	-.8		
$\beta_{23}^{y_9}$	-.7	$\beta_{67}^{y_{10}}$	-.5		
$\beta_{123}^{y_9}$	-.2				

References

Agresti, A., 2002. Categorical data analysis. Wiley, Hoboken.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716-723.

Andrews, R. L., Currim, I. S., 2003. A comparsion of segment retention criteria for finite mixture logit models. Journal of Marketing Research 40, 235-243.

Bouguila, N., ElGuebaly, W., 2009. Discrete data clustering using finite mixture models. Pattern Recognition 42, 33-42.

Bozdogan, H., 1993. Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse-fisher information matrix. In: Opitz, O., Lausen, B., Klar, R. (Eds.), Information and classification, concepts, methods and applications. Springer, Berlin, pp. 40-54.

Breslau, N., Reboussin, B. A., Anthony, J. C., Storr, C. L., 2005. The structure of posttraumatic stress disorder. Archives of General Psychiatry 62, 1343-1351.

Bucholz, K., Hesselbrock, V., Heath, A., Kramer, J., Schuckit, M., 2000. A latent class analysis of antisocial personality disorder symptom data from a multi-centre family stufy of alcoholism. Addiction 95, 553-567.

Collins, L. M., Lanza, S. T., 2010. Latent class and latent transition analysis. Wiley, Hoboken.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society: Series B 39, 1-38.

Dias, J. G., 2006. Model selection for the binary latent class model: A Monte Carlo simulation. In: Batagelj, V., Bock, H. H., Ferligoj, A., Ziberna, A. (Eds.), Data Science and Classification. Springer, Berlin, pp. 91-99.

Efron, B. Tibshirani, R., 1993. An introduction to the bootstrap. Chapman & Hall, London.

Everitt, B. S., Landau, S., Leese, M., 2001. Cluster Analysis. Arnold, London.

Gebregziabher, M. DeSantis, S. M., 2010. Latent class based multiple imputation approach for missing categorical data. Journal of Statistical Planning and Inference 140, 3252-3262.

Goodman, L. A., 1973. The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. Biometrika 60, 179-192.

Goodman, L. A., 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61, 215-231.

Hagenaars, J., McCutcheon, A. (Eds.), 2002. Applied latent class analysis. Cambridge University Press, New York.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177-196.

Keel, P., Fichter, M., Quadflieg, N., Bulik, C., Baxter, M., Thornton, L., et al., 2004. Application of latent class analysis to empirically define eating disorder phenotypes. Archives of General Psychiatry 61, 192-200.

Lazarsfeld, P. F., 1950. The logical and mathematical foundation of latent structure analysis. In: Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., Clausen, J. A. (Eds.), Studies in social psychology in World War II. Vol. IV, Measurement and Prediction. Princeton University Press, Princeton, pp. 361-412.

Lin, T. H., Dayton, C. M., 1997. Model selection information criteria for non-nested latent class models. Journal of Educational and Behavioral Statistics 22, 249-264.

Linzer, D. A., 2011. Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. Political Analysis 19, 173-187.

Little, R. J. A., Rubin, D. B., 2002. Statistical analysis with missing values. Wiley, Hoboken.

Magidson, J, Vermunt, J. K., 2004. Latent class models. In: D. Kaplan (Ed.), Handbook of quantitative methodology for the social sciences. Sage, Newbury Park, pp. 175-198.

McCutcheon, A. L., 1987. Latent Class Analysis. Sage Publications, Newbury Park.

McLachlan, G. J., Peel, D., 2000. Finite mixture models. Wiley, New York.

Mittelhaëuser, M., Van der Ark, L. A., Richards, G. W., 2010. Preparing ATLAS survey data for statistical analysis. Unpublished report. Retrieved February 17, 2012, from http://spitswww.uvt.nl/web/fsw/mto/remapapers/2010MarieAnneMittelhaueserIT3.pdf

Muthén, B., 2004. Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In: D. Kaplan (Ed.), Handbook of quantitative methodology for the social sciences. Sage, Newbury Park, pp. 345-368.

Rubin, D.B., 1987. Multiple imputation for nonresponse in surveys. Wiley, New York.

Schafer, J. L., 1997. Analysis of incomplete multivariate data. Chapman & Hall, London.

Schafer, J. L., Graham, J. W., 2002. Missing data: Our view of the state of the art. Psychological Methods 7, 147-177.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461-464.

Ueda, N., Nakano, R., 2000. EM algorithm with split and merge operations for mixture models. Systems and Computers 31, 930-940.

Van der Ark, L. A., Van der Palm, D. W., Sijtsma, K., 2011. A latent-class approach to estimating test-score reliability. Applied Psychological Measurement 35, 380-392.

Van der Palm, D. W., Van der Ark, L. A., Vermunt, J. K., 2012. A Comparison of Incomplete-Data Methods for Categorical Data. Statistical Methods in Medical Research (submitted for publication).

Vermunt, J. K., Magidson, J., 2008. LG-syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module. Statistical Innovations, Belmont.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., Sijtsma, K., 2008. Multiple imputation of incomplete categorical data using latent class analysis. Sociological Methodology 38, 369-397.

Wang, H. X., Luo, B., Zhang, Q. B., Wei, S., 2004. Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm. Pattern Recognition Letters 25, 1799-1809.

Table 1: AIC and computation time for 13 LC models fitted on the ATLAS survey data

Number of LCs	AIC	Computation time
5	469,422.728	0:04h
10	461,707.592	0:09h
15	458,369.833	0:14h
20	455,712.563	0:19h
25	453,578.838	0:25h
30	452,359.392	0:33h
35	451,258.196	0:39h
40	451,400.102	0:42h
45	449,592.844	0:54h
50	450,231.146	0:56h
55	449,465.083	1:04h
60	448,905.116	1:07h
65	456,395.814	1:12h

Table 2: Pearson's chi-square statistic to quantify the difference between the true frequencies and the estimated frequencies yielded by an HLC model for the marginal tables, $y_9 y_{10}$, $y_8 y_{11}$, and $y_6 y_7 y_{10}$, and the whole data. Three variants of an HLC model (minimum of 0 point increase, 5 point increase, or a decrease in AIC) were crossed with 3 levels of minimum class size (0% , 3%, and 6%). The last column includes the chi-square statistics for the independence model, which can be used as a reference point.

	χ^2									
	HLC-1			HLC-5			HLC-AIC			Indep. Model
Min. N	0	30	60	0	30	60	0	30	60	
$y_9 y_{10}$	.012	.012	.085	.022	.022	.112	.107	.107	.107	88.220
$y_8 y_{11}$	.087	.92	7.293	5.804	5.804	8.983	40.624	40.624	40.624	107.243
$y_6 y_7 y_{10}$	.021	.023	.22	.502	.502	.886	0.918	0.918	0.918	488.052
y	480.57	615.20	1674.71	1692.46	1692.46	1916.30	2813.20	2813.20	2813.20	765088

Table 3: Information on the Variables Used in the Ordinal Regression for the ATLAS Cultural Tourism Research Project 2003 Data (ATLAS, 2004).

Variable		Categories	Number of Missing Values (N = 4292)
I want to find out more about the local culture	1	Totally disagree	154
	2	Disagree	
	3	Neutral	
	4	Agree	
	5	Totally agree	
Gender	1	Male	41
	2	Female	
Age	1	15 or younger	28
	2	16-19	
	3	20-29	
	4	30-39	
	5	40-49	
	6	50-59	
	7	60 or older	
Highest level of educational qualification	1	Primary school	62
	2	Secondary school	
	3	Vocational education	
	4	Bachelor's degree	
	5	Master's or doctoral degree	
Is your current occupation (or former)	1	Yes	149
	2	No	
Admission expenditure	1	0 - < 25 euro	2801
	2	25 - < 50 euro	
	3	50 - < 75 euro	
	4	75 - < 100 euro	
	5	≥ 100 euro	

Table 4: Parameter Estimates and Standard Errors of the First Ordinal Regression for the ATLAS Cultural Tourism Research Project 2003 Data Using Complete Case Analysis, LC Analysis, and HLC Analysis.

	Complete Cases (N = 3950)		LC (K=65)		HLC-1 (min N 1%)	
Predictor	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Gender	-.049	.026	-.052	.026	-.050	.025
Age	-.058	.010	-.062	.009	-.061	.009
Primary School	.000		.000		.000	
Secondary School	-.008	.098	-.039	.092	-.053	.093
Vocational Education	-.080	.098	-.098	.092	-.110	.094
Bachelor's Degree	-.067	.096	-.094	.089	-.105	.091
Master's or doctoral degree	-.091	.097	-.109	.091	-.123	.093
Occupation and culture	-.015	.030	-.017	.030	-.022	.029

Table 5: Parameter Estimates and Standard Errors of the Second Ordinal Regression for the ATLAS Cultural Tourism Research Project 2003 Data Using Complete Case Analysis, LC Analysis, and HLC Analysis.

	Complete Cases (N = 1424)		LC (K=65)		HLC-1 (min N 1%)	
Predictor	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Gender	-.077	.044	-.052	.026	-.050	.025
Age	-.082	.017	-.063	.009	-.061	.009
Primary School	.000		.000		.000	
Secondary School	-.110	.180	-.042	.092	-.058	.093
Vocational	-.152	.181	-.101	.092	-.114	.093
Bachelor's Degree	-.106	.176	-.097	.089	-.109	.091
Master's or doctoral	-.244	.179	-.113	.091	-.127	.093
Occupation and	-.041	.049	-.017	.030	-.020	.029
Admission	.013	.014	.007	.012	.010	.012

Table 6: Data log-likelihood yielded by an LC, HLC (min LL=1, min N of 3%), HLC (min LL=1, min N of 2%), and HLC (min LL=1, min N of 1%) model for the ATLAS data.

Method	Log-likelihood	Computation time
LC (K=65)	-216,043.09	8h12m
HLC(min N of 3%, K=62)	-217,990.25	0h47m
HLC(min N of 2%, K=95)	-213,337.94	0h58m
HLC(min N of 1%, K=149)	-205,340.37	1h06m

Figure 1: An exemplary HLC model with 10 LCs, 6 are end-points (indicated by LCs printed in bold-face) and constitute the density estimate. The explanation of the notation within the LCs is deferred to the HLC section.

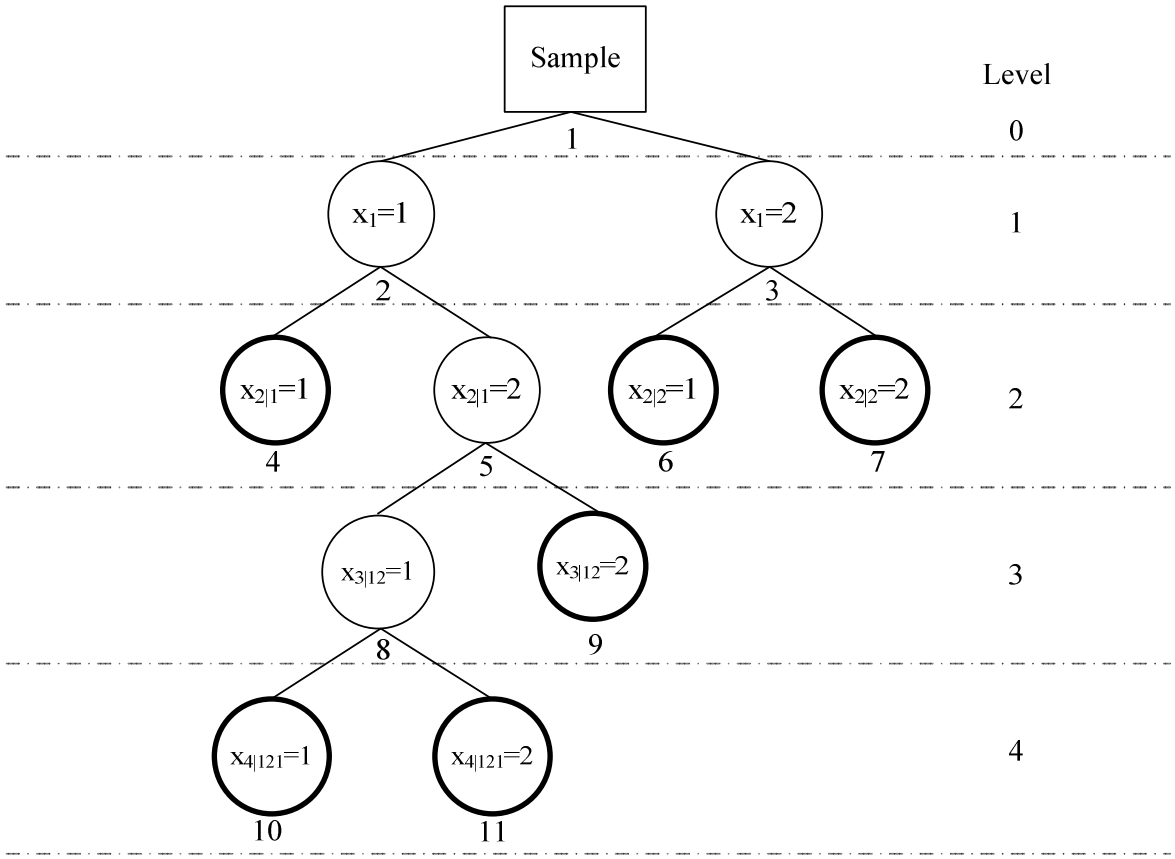

*Figure 2: The three possible situations that can occur when fitting an HLC model; the first split of the data, a second split, and the r*th *split. Each LC has two subscripts: the level at which the LC is located, and a vector of its parent LCs, respectively.*

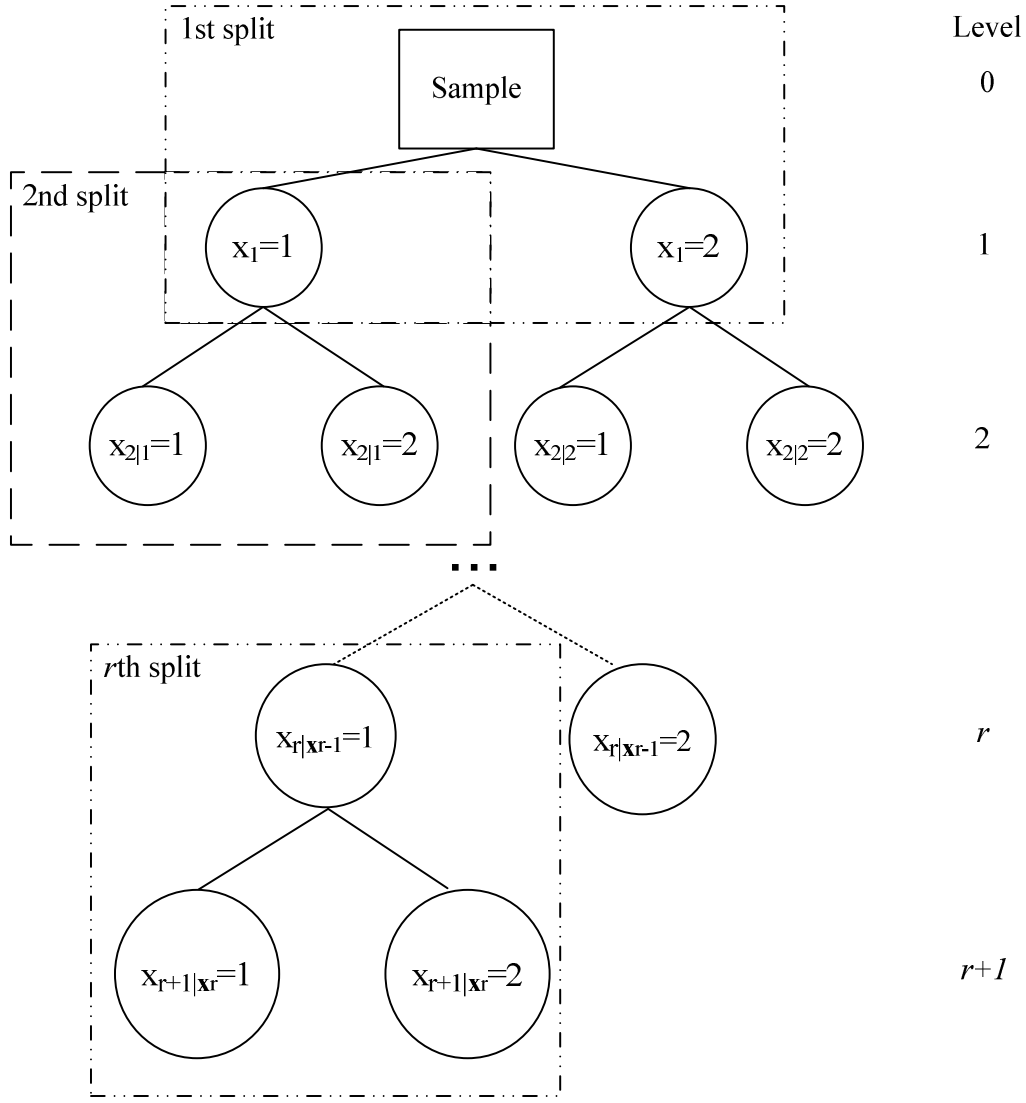

Figure 3: An exemplary HLC analysis to illustrate how the posterior membership probability can be obtained for the r*th split using the local posterior membership probabilities.*

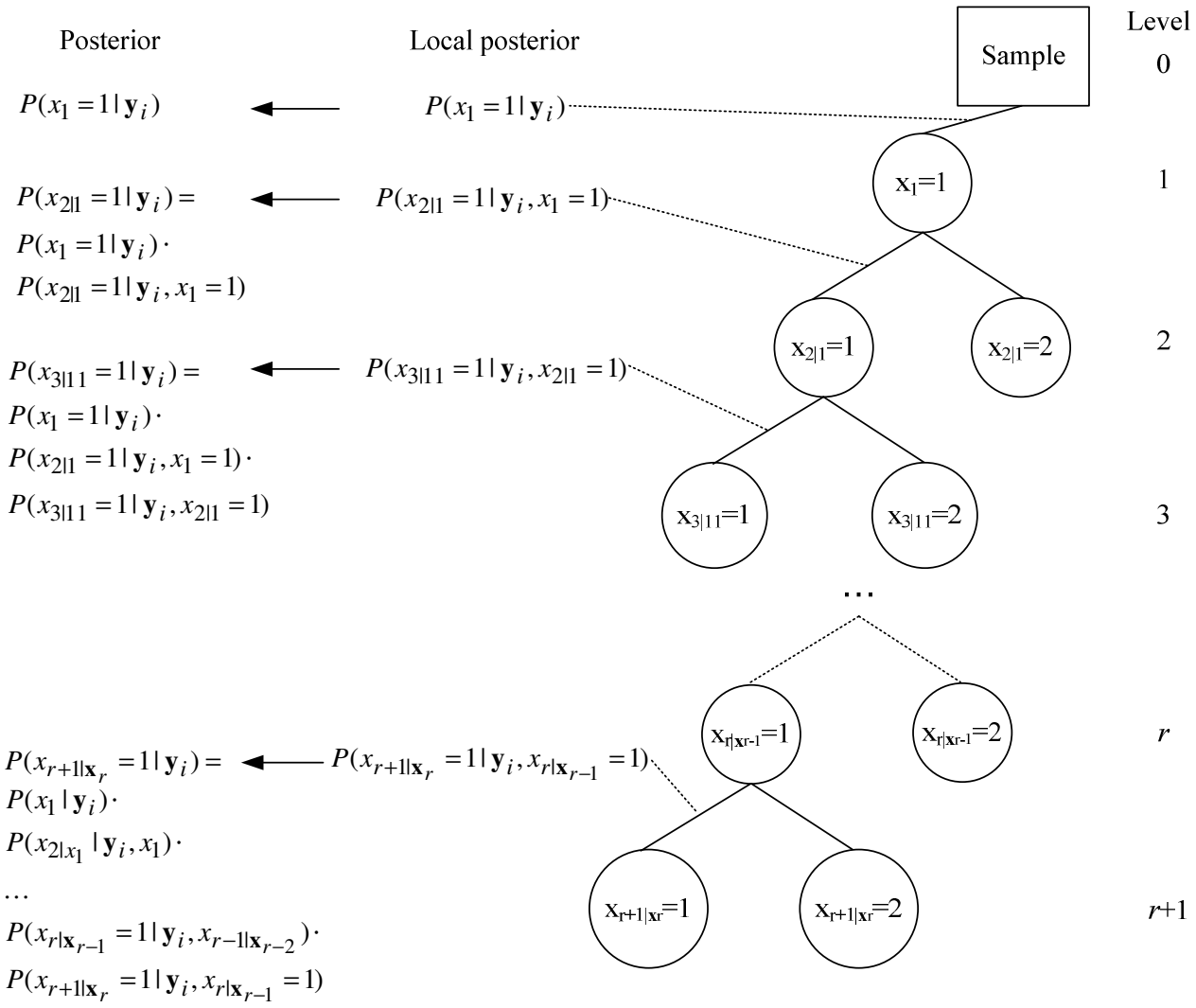

Figure 4: Population model of the generated data example, containing 8 independent ($x_1, \ldots, x_8$) and 3 dependent (y_1, y_2, y_3) dichotomous variables.