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

van der Palm, D.W.; van der Ark, L.A.; Vermunt, J.K.

DOI
10.1007/s00357-016-9195-5

Publication date
2016

Document Version
Final published version

Published in
Journal of Classification

Journal of Classification 33:52-72 (2016)
DOI: 10.1007/s00357-016-9195-5

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

Daniël W. van der Palm

Cito Arnhem, The Netherlands

L. Andries van der Ark

University of Amsterdam, The Netherlands

Jeroen K. Vermunt

Tilburg University, The Netherlands

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 divisive LC (DLC) 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 considerable number of increasingly large LC models may have to be estimated before sufficient model-fit is achieved. A DLC model consists of a sequence of small LC models. Therefore, a DLC model can be estimated much faster and can easily utilize multiple processor cores, meaning that this model is more widely applicable and practical. In this study we describe the algorithm of fitting a DLC model, and discuss the various settings that indirectly influence the precision of a DLC model as a density estimation tool. These settings are illustrated using a synthetic data example, and the best performing algorithm is applied to a real-data example. The generated data example showed that, using specific decision rules, a DLC model is able to correctly model complex associations amongst categorical variables.

Keywords: Latent class analysis; Categorical data; Mixture model; Density estimation; Divisive latent class model; Missing data; Multiple imputation.

Corresponding Author's Address: D.W. van der Palm, Cito, Amsterdamseweg 13 6814 CM Arnhem, The Netherlands, email: danielwillem85@gmail.com.

Published online: 23 February 2016

1. Introduction

Traditionally, latent class (LC) analysis (Lazarsfeld 1950; also see, e.g. Collins and Lanza 2010; Goodman 1974; Hagenaars and McCutcheon 2002; Magidson and Vermunt 2004; McCutcheon 1987; Rindskopf and Rindskopf 1986) is used as a statistical method to identify substantively gmeaningful groups from multivariate categorical data. For example, Keel et al. (2004) distinguished 4 LCs of people with eating disorders that were labeled 'restricting anorexia nervosa', 'anorexia nervosa and bulimia nervosa with the use of multiple methods of purging', 'restricting anorexia nervosa without obsessive-compulsive features', and 'bulimia nervosa with self-induced vomiting as the sole form of purging'. To facilitate interpretation, it is desirable to keep the number of LCs small, and because the interpretation of the LCs is based on the estimated model parameters, it is also desirable that 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. For example, density estimation by means of an LC model has been used for multiple imputation of categorical data (Gebregziabher and DeSantis 2010; Van der Palm, Van der Ark, and Vermunt in press; Vermunt, Van Ginkel, Van der Ark, and Sijtsma 2008), for smoothing large sparse contingency tables (Linzer 2011), for estimating test-score reliability (Van der Ark, Van der Palm, and Sijtsma 2011), and for summarizing image-data bases for pattern recognition (Bouguila and ElGuebaly 2009). 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 among the variables should be accurately described by the LC model. In this context, the LC model is solely used as a tool, and the substantive interpretation of the LCs is not important. Consequently, for density estimation, issues such as model identification, convergence to the global maximum, and having as few LCs as possible do not play a dominant role.

For datasets containing a large number of variables, density estimation by means of an LC model is problematic because a large number of LCs is usually required for precise density estimation. Let LC(K) denote an LC model with K classes. For example, in the context of

handling missing data, Vermunt, Van Ginkel, Van der Ark, and Sijtsma (2008), used AIC (Akaike 1974) as a criterion and selected LC(50) to estimate the joint density of the 79 variables of a survey dataset. They indicated that even more LCs may have been needed for precise density estimation. A typical model-fitting strategy is to estimate LC(5), LC(10), LC(15), LC(20), 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. (2008), containing 4292 cases and 79 categorical variables, and estimated LC(5), LC(10), LC(15), …, LC(60), and LC(65). The analysis took 8 hours and 18 minutes (details in Figure 1, left-hand panel) on a, for current standards, very fast personal computer (i7 2600 quadcore processor, 8GB of internal memory). LC(60) had the lowest AIC value (Figure 1, right-hand panel) and may be taken as the final solution. 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).

As a solution, we introduce the divisive LC (DLC) model as a fast alternative to the LC model for density estimation. First, we provide an intuitive description of the DLC model. Second, we discuss estimation of the DLC model and some arbitrary choices that can be made in the 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. Divisive Latent Class Model

The DLC model is a top-down clustering of respondents into LCs. It is obtained by fitting a sequence of LC(1) and LC(2) models. Figure 2 shows a graphic representation of the structure of a DLC model. It has different levels. In general, let r denote the level in the sequential structure ($r = 0, 1, 2, \ldots$). Each level has a discrete latent variable denoted $X^{(r)}$ ($r = 0, 1, 2, \ldots$). Latent variable $X^{(r)}$ has $Q^{(r)}$ categories, which are the LCs. We use q and s to index the categories of $X^{(r)}$, and we write $X^{(r)} = q$ to denote the qth LC at Level r. This notation is convenient for the formal description of the DLC model. First, at Level 0, we start with LC(1) to describe the joint density of the manifest variables. Hence, X^0 has only one LC ($Q^{(0)} = 1$). In Figure 2, a rectangle is used at Level 0 to indicate that X^0 is in fact a constant rather than a latent variable because the entire sample belongs to the same LC. A decision is made whether or not the goodness of fit would be improved if the LC is split into two LCs. If LC(2) fits

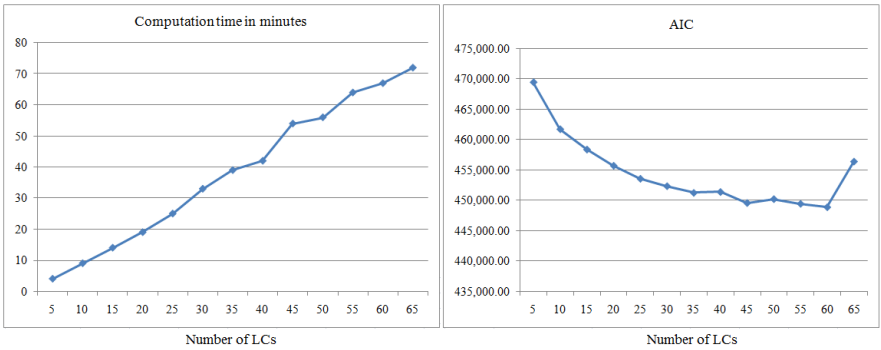

Figure 1. Computation time in minutes (left-hand panel) and AIC value (right-hand panel) for 13 estimated LC Models having, 5, 10, 15, … , 65 latent classes. Each model was fitted on a survey data set consisting of 79 variables.

better than LC(1), then we have two LCs at Level 1, otherwise we have one LC at Level 1 and the procedure stops. Suppose that LC(2) fits better than LC(1) and we have two LCs at Level 1 (case depicted in Figure 2), then for each LC a decision is made whether or not the goodness of fit would be improved by splitting the LC again into two LCs. In Figure 2, the first LC is split whereas the second LC is not, yielding three LCs at Level 2. Suppose that the first LC has weight 0.6; $P(X^{(1)} = 1) = 0.6$, and the second LC has weight 0.4; $P(X^{(1)} = 2) = 0.4$. The splitting of the first LC means that at Level 2 the weight of 0.6 is redistributed across LCs $X^{(2)} = 1$ and $X^{(2)} = 2$. The fact that the second LC is not split implies that at Level 2, $P(X^{(2)} = 3) = 0.4$. Once it has been decided that splitting an LC does not improve the goodness of fit, the particular LC remains unchanged for the rest of the procedure. The splitting procedure continues until splitting LCs no longer improves the goodness of fit. In Figure 2, this is the case at Level 5, where we have six LCs. Numbering the LCs per level from 1 to $Q^{(r)}$is arbitrary. We used the following procedure: Once all LCs at Level r have been either split or maintained, Level r + 1 has been established, and the LCs at Level r + 1 are simply numbered from 1 to $Q^{(r)}$ (from left to right in Figure 2). The DLC model is somewhat similar to divisive clustering, from which we took its name. The difference is that in a DLC model each respondent, at each level, has a probability to belong to each LC (soft partitioning), and in divisive clustering each respondent, at each level, belongs to a cluster with certainty (hard partitioning). The DLC model was inspired by the work of Ueda and Nakano (2000) and Wang, Luo, Zhang, and Wei (2004). Ueda and Nakano introduced a split-and-merge approach to estimating mixture models to overcome the problem of local maxima, whereas Wang et al. used a stepwise split-and-merge

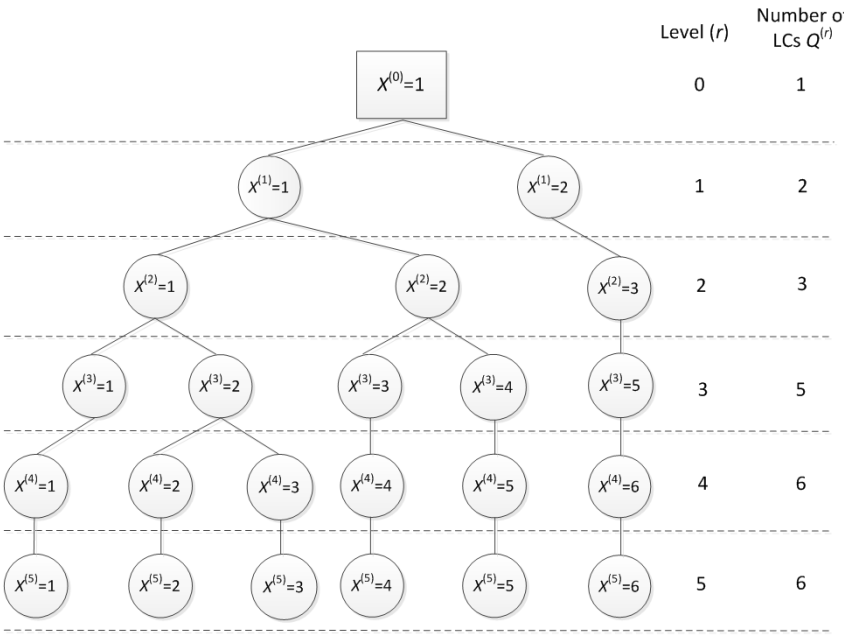

Figure 2. A graphic illustration of the DLC model. For details, see text.

approach to determine the number of components of a mixture model. In the Discussion we compare the DLC model to these papers and related work.

The computational advantage of the DLC model over the standard LC model is that the estimation problem is broken down into a series of small problems coined local problems. Each local problem concerns the question whether splitting LC q at Level r will improve model fit (Figure 3). To this end, at Level $r + 1$, we estimate an LC*(1) model and an LC*(2) model. The asterisks indicate that the models are fitted on a weighted sample in LC q at Level r instead of the unweighted total sample. If the LC*(2) model has a sufficiently better fit than the LC*(1) model, then LC q at Level r will be split. The estimation of the LC*(1) model and the LC*(2) model does not affect the LCs that are not part of the local problem. In the local problem, we arbitrarily number the LCs 1 and 2 for the LC*(2) model and 1 for the LC*(1) model. Note that an LC*(1) model and an LC*(2) model are estimated repeatedly – once for every local problem – in order to investigate whether a split is necessary.

Let X be a discrete latent variable and let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{iJ})$ be the response vector of respondent i to manifest variables $Y_1, \ldots, Y_j, \ldots, Y_J$. In a standard LC($K$) model, the density $P(\mathbf{y}_i)$ is modeled as

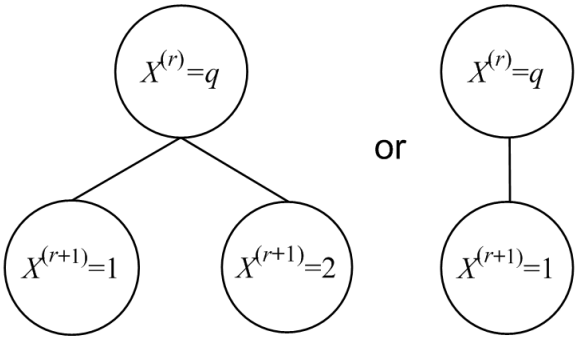

Figure 3. Graphic representation of a local DLC problem: Should an LC at Level r be split into two LCs at Level $r + 1$ (left) or not (right)?

$$P(\mathbf{y}_i; \boldsymbol{\theta}) = \sum_{q=1}^{K} P(X = q) \prod_{j=1}^{J} P(y_{ij}|X = q). \quad (1)$$

The set of parameters, denoted by $\boldsymbol{\theta}$, consists of probabilities $P(X = q)$—the probability that a randomly selected respondent belongs to LC q—and probabilities $P(y_{ij}|X = q)$—the probability that a member of LC q has response y_{ij}. The log-likelihood for the LC(K) model is

$$\log L(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{N} w_i \log \sum_{q=1}^{K} P(X = q) \prod_{j=1}^{J} P(y_{ij}|X = q), \quad (2)$$

where w_i denotes the contribution of the response vector of respondent i to the log-likelihood. For standard LC models, the weights w_i are equal to 1 by definition.

For a local problem, depicted in Figure 3, an LC*(1) model ($K = 1$) and an LC*(2) model ($K = 2$) are estimated for the sample in LC q at Level r. It may be noted that the sample in LC q at Level r consists of the entire sample, in which each observation has been reweighted (to be discussed shortly) rather than a subsample of the observations. Hence, $P^*(y_i) \equiv P(y_i|X^{(r)} = q)$ is modeled rather than $P(y_i)$. We will use q as the index of LCs at Level r, and s as the index of LCs at Level $r + 1$. The LC model in Equation 1, then becomes

$$P^*(\mathbf{y}_i; \boldsymbol{\theta}^*) = \sum_{s=1}^{K} P^*(X^{(r+1)} = s) \prod_{j=1}^{J} P^*(y_{ij}|X^{(r+1)} = s). \quad (3)$$

In Equation 3, density $P^*(y_i)$ is modeled by local parameters; the local parameters have the same interpretation as the parameters of a standard LC model, except for the fact that they are conditional on being member of

LC $X^{(r)} = q$. Thus, $P^*\left(X^{(r+1)} = s\right) \equiv P\left(X^{(r+1)} = s \middle| X^{(r)} = q\right)$ and $P^*\left(y_{ij} \middle| X^{(r+1)} = s\right) \equiv P\left(y_{ij} \middle| X^{(r+1)} = s; X^{(r)} = q\right)$. The local parameters are denoted by $\boldsymbol{\theta}^*$. The subsample in LC q at Level r is fuzzy because each respondent has a probability of belonging to this LC. The probability that a respondent having response vector $\mathbf{y}_i$ belongs to LC q at Level r is denoted as $P(X^{(r)} = q|\mathbf{y}_i)$, and referred to as the *posterior membership probability*. The posterior membership probability determines the weight of a particular respondent in the log-likelihood:

$$w_{iq}^{(r)} = P(X^{(r)} = q|\mathbf{y}_i).$$

Hence, the log-likelihood for the LC*(*K*) model ($K = 1, 2$) in the local problem is

$$\log L(\boldsymbol{\theta}^*; \mathbf{y}) = \sum_{i=1}^{N} w_{iq}^{(r)} \log \sum_{s=1}^{K} P^*\left(X^{(r+1)} = s\right) \prod_{j=1}^{J} P^*\left(y_{ij} \middle| X^{(r+1)} = s\right). \quad (4)$$

The parameter estimates of the selected LC model at Level r+1 in the local problem also yields a *local posterior membership probability* for $X^{(r+1)}$:

$$P^*(X^{(r+1)} = s|\mathbf{y}_i) \equiv P^*(X^{(r+1)} = s \middle| \mathbf{y}_i; X^{(r)} = q) = \frac{P^*\left(X^{(r+1)}=s\right) \prod_{j=1}^{J} P^*\left(y_{ij} \middle| X^{(r+1)}=s\right)}{\sum_{k=1}^{K} P^*\left(X^{(r+1)}=k\right) \prod_{j=1}^{J} P^*\left(y_{ij} \middle| X^{(r+1)}=k\right)}. \quad (5)$$

For the LC* (2) model, the local posterior is the probability that a respondent belongs to each of the two LCs at level r + 1, conditional on being member of LC q at level r. For the LC*(1) model, the local posterior equals 1 by definition.The local posterior membership probability is used to determine the weights of the respondents in the likelihood for the local problems at the next level. Hence, the sample is not physically separated out, but only the weights change. The weights at Level r + 1 are obtained by multiplying the local posterior probability at Level r + 1 and the (global) posterior probability at Level r:

$$w_{is}^{(r+1)} = P\left(X^{(r+1)} = s \middle| \mathbf{y}_i\right) = P^*\left(X^{(r+1)} = s \middle| \mathbf{y}_i\right) \times P\left(X^{(r)} = q \middle| \mathbf{y}_i\right). \quad (6)$$

The DLC model is estimated by the following iterative procedure.

1. *Initial step*: At Level 0, set $w_{i1}^{(0)} := 1$, $P\left(X^{(0)} = 1\right) := 1$, and $Q^{(0)} := 1$.
2. *Solve the local problem of LC q at Level r:* Estimate an LC*(1) and LC*(2) model for the fuzzy sample in LC $X^{(r)} = q$ by optimizing the likelihood in Equation 4, and choose either an LC*(1) model or LC*(2) model. If an LC*(1) model is chosen, LC $X^{(r)} = q$ is no

longer considered for division at later levels and steps 3, 4, and 5 are skipped; see discussion hereunder.

3. *Compute the local posterior membership probabilities* (Equation 5).
4. *Update the posterior membership probabilities* from the local membership probabilities (Equation 6). The updated posterior membership probabilities are the weights for the local problems at Level $r + 1$.
5. *Update the parameter estimates using the posterior membership probabilities and local parameter estimates.*
   - $P\left(X^{(r+1)} = s\right) = \frac{1}{N}\sum_{i=1}^{N} P\left(X^{(r+1)} = s|\boldsymbol{y}_i\right)$
   - $P\left(y_{ij}\middle|X^{(r+1)} = s\right) = P^{*}\left(y_{ij}\middle|X^{(r+1)} = s\right)$
6. Repeat steps 2 through 5 for all LCs at Level r.
7. Renumber the LCs from 1 to $Q^{(r)}$, and let $r = r + 1$.
8. Repeat steps 2 to 7 until no more classes are split.

The remaining problem of DLC estimation is the choice of either the LC*(1) model or the LC*(2) model in each local problem (Figure 3). The choice depends on the required precision and the sample size. If the number of LCs becomes too large, the density estimate may be based on chance capitalization. If the number of LCs becomes too small, the density estimation may not be precise enough. Relevant factors for the choice of either the LC*(1) model or the LC*(2) model may be the difference in log-likelihood, the sample sizes in the LCs, and the size of residual associations between the observed variables. In the generated data study, we investigate this issue.

3. Generated Data Study

The main question was whether the DLC model can precisely estimate a complex density that was not generated by an LC model. To this end, we used a DLC model to estimate a complex density under ideal circumstances, so removing all influences of sampling error. Additionally, we investigated different choices for selecting an LC*(1) model or an LC*(2) model in the local problem.

3.1 Method

We defined a complex population model, depicted in Figure 4, for 11 dichotomous variables ($Y_1, \ldots, Y_{11}$). The population model consists of two sets of independent variables ($\{Y_1, Y_2, Y_3\}$ and $\{Y_4, \ldots, Y_8\}$) and three

dependent variables Y_9, Y_{10}, and Y_{11}. Log-linear models describe the associations among the independent variables, and logit models described the relations between the independent and dependent variables. The model contained several two-way and three-way interactions. The appendix gives the details of the population model. Multiplying the population probability for each of the $2^{11} = 2048$ response patterns by 1,000 produced the frequencies for all response patterns, amounting to a sample (of size $N = 1{,}000$) that is exactly in accordance with the population.

We compared the true and the estimated marginal probabilities by a DLC 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 a DLC model is able to correctly pick up this complex association. The estimated marginal probabilities can be computed from the estimated DLC parameters. For example, the probabilities of Y_6, Y_7, and Y_{10} can be obtained by $\hat{P}(Y_6, Y_7, Y_{10}) = \sum_{s=1}^{K} \hat{P}(X = s)\, \hat{P}(Y_6|X = s)\hat{P}(Y_7|X = s)\hat{P}(Y_{10}|X = s)$. As an outcome variable we reported Pearson's chi-squared statistic for the differences between the true and the estimated expected frequencies for a sample size of $N = 1{,}000$. The degrees of freedom are equal to 3 for the two-way interactions, 7 for the three-way interactions, and 2047 for the entire density.

We investigated 21 decision rules for model selection in the local problem. Each decision rule is a combination of a model-fit criterion (7 levels) and a minimum sample size for an LC (3 levels):

1. *Model fit criteria*. Using a model-fit criterion implies that an LC is split if the resulting LC*(2) model shows better fit than the LC*(1) model. Six model-fit criteria were a combination of a minimum increase in the log-likelihood (levels 'at least 1 point increase' and 'at least 3 points increase'), and a maximum value of the highest standardized bivariate residual (levels 'unrestrictive', 'stop if all bivariate residuals are less than 1', and 'stop if all bivariate residuals are less than 3)'. A 7th model-fit criterion is to keep splitting LCs as long as AIC decreases, which amounts to a minimum improvement of the log-likelihood equal to the number of additional parameters. We refer to Vermunt et al. (2008) for the discussion of preferring AIC over other relative fit statistics, such as AIC3 and BIC, in density estimation.

2. *Minimum sample size for an LC*. Using this criterion means that an LC is only considered for splitting if it contains a minimal number of respondents. We considered minimal sample sizes of 0, 30, and 60.

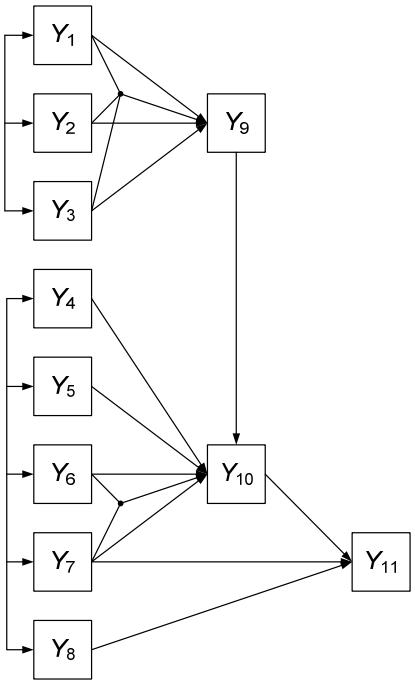

Figure 4. Population model used in the generated-data study. For details, see text.

The DLC model was estimated for all decision rules. The question whether the DLC model is able to accurately describe a complex density under ideal circumstances was investigated by examining the difference in true and estimated marginal probabilities under the least restrictive levels of the decision rules. This cell was used as an upper benchmark for investigating the effect of using more stringent levels of the decision rules. Note that levels of the decision rules of the upper benchmark should not be used in practice because one would also model all sampling fluctuations. Yet, comparing more stringent levels of decision rules to the upper benchmark is useful because it shows the relative decrease of precision in estimating the two-way and three-way interactions. Using a similar train of thought we used the independence model as a lower benchmark.

3.2 Results

For the upper bench mark (Table 1, first row), the values of the chi-squared statistics were very small compared to the degrees of freedom, indicating that the DLC can pick up complex associations in the data under

Table 1. Chi-square statistics for the difference between the estimated frequencies using a DLC model and the true frequencies of three marginal tables: y_9y_{10}, y_8y_{11}, and $y_6y_7y_{10}$, and the total data. seven variants of a DLC model were crossed with three levels of minimum class size (0 , 30, and 60). The first row contains the least restrictive decision rules, and serves as an upper benchmark, the last row contains the chi-square values for the independence model as a lower benchmark.

Decision rules			Marginals			
L	Residual	*N*	y_9y_{10}	y_8y_{11}	$y_6y_7y_{10}$	Total
1	0	0	.012	.065	.021	69.118
		30	.012	.897	.023	81.855
		60	.085	4.400	.208	187.064
1	1	0	.012	.053	.021	81.392
		30	.012	.896	.023	92.402
		60	.085	4.398	.208	192.870
1	5	0	.022	3.033	.470	164.787
		30	.022	3.033	.470	167.080
		60	.111	8.498	.828	219.414
5	0	0	.011	5.718	.059	174.113
		30	.011	5.718	.059	174.113
		60	.111	8.498	.828	219.414
5	1	0	.022	5.541	.477	180.048
		30	.022	5.541	.477	180.048
		60	.111	8.498	.828	219.414
5	5	0	.022	5.541	.477	180.048
		30	.022	5.541	.477	180.048
		60	.111	8.498	.828	219.414
AIC	0	0	.107	36.988	.857	337.374
		30	.010	37.639	.064	342.344
		60	.107	36.862	.858	331.970
Independence model			80.661	96.097	374.989	5740.297

ideal circumstances. Note that these values are too good to be true because the data contain no error, and an additional simulation study (not tabulated) using data that were sampled from the model showed that the chi-square statistic – quantifying the difference between the estimated density and the population density – may actually increase if too many divisions are made

(i.e. overfitting the data). As expected the lower benchmark (Table 1, last row) showed high values of the chi-squared statistics compared to the degrees of freedom, indicating that the independence model cannot describe complex associations in the data.

The more conservative decision rules especially deteriorated the model-fit for the two-way interaction of Y_8 and Y_{11}. The additional safeguards to have at least 30 respondents in each cell and to stop splitting when all standardized residuals are less than 1 did not affect the precision greatly. Other levels of the decision rules seriously deteriorated the density estimate, in particular choosing AIC as a criterion seems insufficient.

4. Real-Data Example

A well-known application of the LC model as a density estimator is to handle missing data problems by means of multiple imputation (e.g. Vermunt et al. 2008; Gebgziabher and DeSantis 2010; Van der Palm, Van der Ark and Vermunt 2012). The procedure consists of the following steps. First, m nonparametric bootstrap samples are drawn from the incomplete data. Typically $m = 5$, but larger values may improve the quality of the statistical analysis. Second, for each bootstrap sample, a well-fitting LC model is estimated, yielding m LC-models numbered $1, \dots, m$. Third, m completed data sets are constructed, where completed data set i ($i = 1, \dots, m$) was constructed by replacing the missing values of the original incomplete data by a value drawn from model i. Fourth, the statistical analysis of choice is performed m times, on each completed dataset. Finally, the parameters of interest in the m analyses are combined using Rubin's (1987) rules. For details of applying the LC model in the context of missing data, we refer to Vermunt et al. (2008).

We applied this procedure and used both the LC model and the DLC model to estimate the joint density (i.e., second step of the procedure) to illustrate the effect on the computation time of the entire procedure and the effect on the parameter estimates at the end of the procedure. The LC model yielding the lowest AIC value was selected, and for DLC we used a minimum increase of the log-likelihood of 1 point and a minimal sample size of 30 (see Table 1, second row) as a criterion for splitting.

We expected that using the DLC model is much faster than using the LC model, and we expected that the effect on the parameter estimates and their standard errors would be negligible. The latter indicates that the DLC model yields a sufficiently precise density estimate. To be able to assess whether the effect on the parameter estimates is small, we also used complete-case analysis as a benchmark. Complete-case analysis may have a large effect on the parameter estimates (e.g. Schafer and Graham 2002).

We analyzed a dataset from the ATLAS Cultural Tourism Research Project (Richards 2010), a survey that addresses topics such as motivations, activities, and impressions of visitors of cultural sites and events. The dataset contained the scores of 4292 respondents on 79 categorical variables: 52 had two categories, one had three categories, 19 had five categories, two had six categories, and the remaining five variables had 7, 8, 9, 10 and 17 categories, respectively. Complete information was available for only 794 respondents. Steps 1, 2, and 3 of the procedure were carried out as described above performing with either $m = 10$ LC analyses or $m = 10$ DLC analyses on all 79 variables in step 2.

In step 4, we used two (adjacent-category) ordinal regression models (Agresti 2002, pp. 286–288) to predict the responses to the question "I want to find out more about the local culture" with four and five explanatory variables, respectively. In the first model, the variable Admission Expenditure was excluded as an explanatory variable, rendering 3950 complete cases, whereas in the second model this variable was included, which reduced the number of complete cases to 1424. Table 2 shows the details of the variables involved.

The LC(65) model and the DLC(95) model were selected. Table 3 shows the coefficients and standard errors of the two ordinal regression models, estimated after complete-case analysis, multiple impution using the LC(65) model, and multiple imputation using the DLC(95) model. For the first regression model (Table 3, upper panel), Age and Gender had a negative effect, which means that younger people and men have a greater desire to learn more about local culture. The effect of the other explanatory variables was not significant. The parameter estimates and standard errors are rather similar across the methods, except for some differences in the parameter estimates for Education. Because there is only a small proportion of missing values in this analysis, it is not surprising that complete-case analysis and MI gave similar results. It is reassuring that for most regression coefficients, MI using the LC model and the DLC model provided similar estimates. For the second regression analysis (Table 3, lower panel), the substantive results did not change because the effect of the added explanatory variable Admission Expenditure was not significant, but we found significant differences in parameter estimates between complete-case analysis and MI. The estimates based on MI are similar in the two regressions, whereas the estimates based on complete-case analysis have changed: The estimated coefficients of age, gender, and education nearly doubled and all standard errors became larger. Admission Expenditure had no significant effect, but its inclusion resulted in more cases having missing values and consequently the effects of Age and Gender were overestimated when complete-case analysis was used.

Table 2. Variables used in the ordinal regression for the ATLAS Cultural Tourism Research Project 2003 data.

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	
Is your current occupation (or former) connected with culture?	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 3. Parameter estimates and standard errors of two ordinal regression for the ATLAS Cultural Tourism Research Project 2003 data using Complete Case Analysis, MI using an LC model, and MI using a DLC model (minimum increase of the log-likelihood of 1 point and a minimal sample size of 30).

			Multiple imputation			
	Complete-case Analysis		LC (K=65)		DLC (95)	
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	-.054	.093
Vocational Education	-.080	.098	-.098	.092	-.110	.094
Bachelor's Degree	-.067	.096	-.094	.089	-.105	.091
Master's or doctoral	-.091	.097	-.109	.091	-.124	.093
Occupation and culture	-.015	.030	-.017	.030	-.021	.029

			Multiple imputation			
	Complete Cases (N = 1424)		LC (K=65)		DLC (95)	
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	-.106	.176	-.097	.089	-.109	.091
Master's or	-.244	.179	-.113	.091	-.128	.093
Occupation and	-.041	.049	-.017	.030	-.021	.029
Admission	.013	.014	.007	.012	.010	.012

Table 4. Data log-likelihood and computation time under four different models.

Method	Log-likelihood	Computation time
LC (65)	-216,043.09	8h12m
DLC(62)[1]	-217,990.25	0h47m
DLC(95)[2]	-213,337.94	1h02m
DLC(149)[3]	-205,340.37	1h06m

Note: 1 = minimum class-size at least $N = 60$, 2 = minimum class-size at least $N = 30$, 3 = minimum class-size at least $N = 10$.

Although we cannot compare the estimates of the three methods to the population values, this result indicates that both the LC model and the DLC model perform well in this application. It is reassuring that the results based on the LC model and the DLC model are similar and largely concur with those of complete-case analysis when the proportion of missingness is small and that the estimates are stable across the two regression analyses.

Table 4 shows the log-likelihood and the computation time for the LC model and the DLC model used for MI in the real-data example, plus for some alternative DLC models. The computation time for the standard LC model also includes the required computation time to estimate the LC models with fewer LCs. Table 4 shows that the DLC (minimal improvement in the LL of 1 point and minimal sample size of 30) and DLC-1 (minimal improvement in the LL of 1 point and minimal sample size of 10) models yield a better fit than the LC(65) model, and in much less time.

5. Discussion

For density estimation, the DLC had three advantages over the standard LC model. First, in the processes of finding a well-fitting LC model, say LC(K), standard LC analysis requires estimating K models, whereas DLC analysis requires estimating one model. Hence, it is no longer necessary to manually estimate and compare several models. Second, in standard LC analysis, the number of LCs is specified a priori, whereas in DLC analysis it is not; the number of LCs is increased during the estimation process until a sufficiently precise density estimate is obtained. Third, each LC model starts from scratch: the information in an LC(K) model is neglected when fitting an LC($K + 1$) model, whereas the DLC model is a sequence of small local problems and each local problem

at Level $r + 1$ takes into account the information obtained at Level r. Due to this efficiency and relative simplicity, DLC estimation is much faster than LC estimation.

Several authors (e.g. Ueda and Nakano 2000; Wang, Luo, Zhang, and Wei 2004; Hoijtink and Notenboom 2004) also have proposed a splitting procedure. We would like to stress that these approaches serve different purposes. The procedures proposed by Ueda and Nakano and Wang et al. pertain to continuous data whereas our procedure pertains to discrete data. In addition, they use the splitting only to obtain starting values. For instance, in case of a mixture of Q normal distributions, they split one of the classes yielding a mixture of $Q + 1$ normal distributions, and they use this solution as starting values for an unrestricted mixture of $Q + 1$ normal distributions. Hoijtink and Notenboom (also, see Van Hattum and Hoijtink 2009) used the splitting of LCs as a clever trick enabling the estimation of a traditional LC model by means of a Gibbs sampler: They split the largest LC (and if that does not work the second largest LC, etc.) and also use the solution as starting values for an unrestricted LC($Q + 1$) model. Without the splitting, the Gibbs sampler would not be able to estimate models with large numbers of classes. Our approach is substantially different because it produces a truly divisive solution: The relation between the levels is known, whereas Ueda and Nakano (2000), Wang, Luo, Zhang and Wei (2004) and Hoijtink and Notenboom (2004) do not use such levels; their solution produces a series of mixtures with 1, 2, 3 ... LCs, just as an ordinary LC analysis.

The generated data example showed that the DLC model is able to pick up two-way and three-way associations from a complex population model. The suggested decision rules for splitting classes worked well for the generated-data example. In a missing-data context, Vermunt et al. (2008) found that over-fitting did not pose a big problem when using an LC model for density estimation suggesting rather liberal stopping rules. However, Van der Palm, Van der Ark, and Sijtsma (2014) found that density estimation using the DLC model in the context of test-score reliability required a conservative stopping criterion. Hence, a systematic evaluation of the DLC density estimation procedure is required, including the effect of different stopping rules and the effect of sample size. This is a topic for future research. As an aside, in a Bayesian framework, Richardson and Green (1997) introduced the birth-and-death algorithm to automatically determine number of components for mixture models, and it would be interesting to investigate whether such a Bayesian methodology would work well for density estimation.

The real-data example showed that a DLC model can easily be applied to a dataset with a large number of cases and polytomous variables. For a standard LC model with 65 LCs it took more than 8 hours to

establish the best fitting model for this dataset, whereas a DLC model only required 1 hour and 2 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. The application that we discussed underlines the benefits of a DLC model. If a researcher wants to use MI, the density of the data has to be estimated several times (10 times in this case). Hence, using a DLC model for MI instead of an LC model reduced the runtime for this dataset from 83h (10*8h18m) to 10h20m (10*1h2m).

DLC estimation has now been implemented in the software package LatentGOLD (Vermunt and Magidson 2008) which makes it easier to apply the method. As an aside, we note that it is relatively easy to use multiple processing cores for the estimation of a DLC model because estimating the DLC model boils down to estimating a sequence of independent local problems. For the standard LC model, the processing load would have to be divided and delegated to each processor core within one estimation algorithm, which is more difficult and less efficient. For example, suppose a computer has four processor cores. After the first split (e.g. Figure 2), one processor cores can handle the estimation of the LCs beyond the first LC, and a second processing core can handle the estimation of the LCs beyond the second LC. After another split, the third processor core can be used. This makes the estimation process even faster.

Appendix

The densities are described in terms of the realizations of Y, denoted by y. Let β_j denote a log-linear parameter value. The joint density of y_1, y_2, and y_3 is defined as,

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

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-variables associations. Table A1 shows the actual values of the parameters.

The joint density of $y_4, y_5, y_6, y_7,$ and y_8 is defined as

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

and only contains two-way associations. Table A2 shows the actual values of the parameters.

Table A1. Log-linear parameters for the density of Y_1, Y_2, and Y_3.

Parameter	Value
β_1^1	.2
β_2^1	-.6
β_3^1	.4
β_4^1	.2
β_5^1	.6
β_{23}^2	-.1
β_{123}^3	.2

Table A2. 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_4^1	.2	β_{45}^2	.4	β_{57}^2	.6
β_5^1	-.6	β_{46}^2	-.2	β_{58}^2	-.2
β_6^1	.4	β_{47}^2	.6	β_{67}^2	.1
β_7^1	.2	β_{48}^2	-.3	β_{68}^2	-.2
β_8^1	.6	β_{56}^2	.8	β_{78}^2	.6

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 jth independent variable. For dependent variable y_9,

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

for dependent variable y_{10},

$$\text{logit}(y_{10}) = \beta_0^{y_{10}} + \beta_9^{y_{10}} y_9 + \beta_4^{y_{10}} y_4 + \beta_5^{y_{10}} y_5 + \beta_6^{y_{10}} y_6 + \beta_7^{y_{10}} y_7 + \beta_8^{y_{10}} y_8 + \beta_{78}^{y_{10}} y_7 y_8,$$

and for dependent variable y_{11},

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

These relationships yield a complex density including three-way associations. Table A3 shows the values of the logistic regression parameters.

Table A3. 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*, New York: Wiley.

AKAIKE, H. (1974), "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control, 19,* 716–723.

BOUGUILA, N., and ELGUEBALY, W. (2009), "Discrete Data Clustering Using Finite Mixture Models", *Pattern Recognition, 42,* 33–42.

COLLINS, L.M., and LANZA, S.T. (2010), *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*, Hoboken, NJ: Wiley.

EVERITT, B.S., LANDAU, S., and LEESE, M. (2001), *Cluster Analysis,* London: Arnold.

GEBREGZIABHER, M., and 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. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models", *Biometrika, 61,* 215–231.

HAGENAARS, J.A., and MCCUTCHEON, A.L. (eds.) (2002), *Applied Latent Class Analysis*, Cambridge, UK: Cambridge University Press.

HOIJTINK, H., and NOTENBOOM, A. (2004), "Model Based Clustering of Large Data Sets: Tracing the Development of Spelling Ability", *Psychometrika, 69,* 481–498.

KEEL, P., FICHTER, M., QUADFLIEG, N., BULIK, C., BAXTER, M., THORNTON, L., et al. (2004), "Application of Latent Class Analysis to Empirically Defined Eating Disorder Phenotypes", *Archives of General Psychiatry, 61,* 192–200.

LAZARSFELD, P.F. (1950), "The Logical and Mathematical Foundation of Latent Structure Analysis", in *Studies in Social Psychology in World War II. Vol. IV: Measurement and Prediction*, eds. S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star and J.A. Clausen, Princeton, NJ: Princeton University Press, pp. 361–412.

LINZER, D.A. (2011), "Reliable Inference in Highly Stratified Contingency Tables: Using Latent Class Models as Density Estimators", *Political Analysis, 19*, 173–187.

MAGIDSON, J., and VERMUNT, J.K., (2004), "Latent Class Models", in *Handbook of Quantitative Methodology for the Social Sciences,* ed. D. Kaplan, Newbury Park, NJ: Sage, pp. 175–198.

MCCUTCHEON, A.L. (1987), *Latent Class Analysis,* Newbury Park, CA: Sage.

MCLACHLAN, G.J., and PEEL, D. (2000), *Finite Mixture Models,* New York: Wiley.

RICHARDS, G. (2010), "The Traditional Quantitative Approach Surveying Cultural Tourists: Lessons From the ATLAS Cultural Tourism Research Project", in C*ultural Tourism Research Methods,* eds. G. Richards and W. Munsters, Wallingford, UK: CABI, pp. 13–32.

RICHARDSON, S., and GREEN, P.J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components", *Journal of the Royal Statistical Society. Series B, 59*, 731–792.

RINDSKOPF, D., and RINDSKOPF, W. (1986), "The Value of Latent Class Analysis in Medical Diagnosis", *Statistics in Medicine, 5,* 21–27.

RUBIN, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys,* New York: Wiley.

SCHAFER J.L., and GRAHAM J.W. (2002), "Missing Data: Our View of the State of the Art", *Psychological Methods, 7,* 147–177.

UEDA, N., and 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., and 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., and SIJTSMA, K. (2014), "A Flexible Latent-Class Approach to Estimating Test-Score Reliability", *Journal of Educational Measurement*, *51*, 339–357.

VAN DER PALM, D.W., VAN DER ARK, L.A., and VERMUNT, J.K. (Advance Online Publication), "A Comparison of Incomplete-Data Methods for Categorical Data", *Statistical Methods in Medical Research*, doi: 10.1177/0962280212465502, http://smm.sagepub.com/content/early/2012/11/15/0962280212465502.abstract.

VAN HATTUM, P., and HOIJTINK, H. (2009), "Market Segmentation Using Brand Strategy Research: Bayesian Inference With Respect to Mixtures of Log-Linear Models", *Journal of Classification, 26,* 297–328.

VERMUNT J.K., and MAGIDSON J. (2008), *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*, Belmont, MA: Statistical Innovations.

VERMUNT, J.K., VAN GINKEL, J.R., VAN DER ARK, L.A., and 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., and 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.