Latent Class Analysis

Jeroen K. Vermunt,

Department of Methodology and Statistics,

Tilburg University,

PO Box 90153,

5000LE Tilburg,

The Netherlands

Email: j.k.vermunt@uvt.nl

Version of January 12 2022

Will be published as:

Vermunt, J.K. (2022). Latent class analysis. In: P. Peterson, E. Baker, B. McGaw, F. Rizvi, G. Smith, and K. Gutierrez (eds.), *International Encyclopedia of Education, Fourth Edition,* Oxford: Elsevier.

Latent Class Analysis

Jeroen K. Vermunt, Tilburg University, Tilburg, The Netherlands

Keywords:

latent class analysis, latent profile models, mixture model, finite mixture model, random effects modeling, scaling models, cluster analysis, latent Markov models, statistical software, mixture regression, mixture growth models, three-step analysis

Abstract:

A statistical model can be called a latent class (LC) or mixture model if it assumes that some of its parameters differ across unobserved subgroups, latent classes, or mixture components. This rather general idea has several seemingly unrelated applications, the most important of which are clustering, scaling, density estimation, and random-effects modeling. This article describes simple LC models for clustering, restricted LC models for scaling, and mixture regression models for nonparametric random-effects modeling, as well as gives an overview of recent developments in the field of LC analysis. Moreover, attention is paid to topics such as maximum likelihood estimation, identification issues, model selection, and software.

Latent Class Analysis

Jeroen K. Vermunt, Tilburg University, Tilburg, The Netherlands

Introduction

A statistical model can be called a latent class (LC) or mixture model if it assumes that some of its parameters differ across unobserved subgroups, latent classes, or mixture components. This rather general idea has several seemingly unrelated applications, the most important of which are clustering, scaling, density estimation, and random-effects modeling. It should be noted that in applied fields the terms LC model and mixture model are often used interchangeably, which is also what I will do here. In the more technical statistical literature on mixture modeling, the term LC analysis is reserved for a specific type of mixture model, that is, a mixture model for a set of categorical items (for the classical LC model).

LC analysis was introduced in 1950 by Lazarsfeld as a tool for building typologies (or clustering) based on dichotomous observed variables. More than 20 years later, Goodman (1974) made this model applicable in practice by developing an algorithm for obtaining maximum likelihood estimates of the model parameters, as well as proposed extensions for polytomous manifest variables and did important work on the issue of model identification. Many important extensions of this classical LC model have been proposed since then,

such as models containing explanatory variables (Dayton and Macready, 1988), models that relax the local independence assumption (Hagenaars, 1988), constrained models similar to IRT models (Lindsay, Clogg and Grego, 1991; Heinen, 1996), models with multiple latent variables (Magidson and Vermunt, 2001), models for longitudinal data (Van de Pol and Langeheine, 1990; Collins and Lanza, 2010), models for multilevel data (Vermunt, 2003, 2010b), three-step LC analysis (Bolck, Croon, and Hagenaars, 2004; Vermunt, 2010a), and LC tree models (Van den Bergh, van Kollenburg, and Vermunt, 2018).

Whereas the classical LC model and its extensions are conceived primarily as a clustering and scaling tool for categorical data analysis, LC and finite mixture models can be useful in several other areas as well. One of these is as a probabilistic cluster analysis tool for continuous observed variables, an approach that offers many advantages over traditional cluster techniques such as K-means clustering (Wolfe, 1970; McLachlan and Peel, 2000; Vermunt and Magidson, 2002). Another application area is dealing with unobserved heterogeneity, as happens in mixture regression analysis of multilevel or repeated measurement data (Wedel and De Sarbo, 1994; Vermunt and Van Dijk, 2001).

The remainder of this article is organized as follows. After introducing the simplest type of LC models, I discuss various restricted LC models as well as models with explanatory variables. Next, I give an overview of other types of LC and mixture models, which includes various recently proposed extensions. The

remaining sections focus on parameter estimation, model selection, and software. An annotated list of further reading is provided at the end of the chapter.

Simple LC and mixture models

LC analysis is typically used as a tool for analyzing multivariate response data; that is, data consisting of several dependent variables, response variables, or items. We will denote the response of subject *i* on dependent variable *j* by y_{ij}, and the number of dependent variables by *J*. The full response vector of a subject is denoted by $\mathbf{y}_i$. To make things more concrete, Table 1 presents a small illustrative data set consisting of three dichotomous responses, y_{i1}, y_{i2}, and y_{i3} (0=incorrect; 1=correct). This is a subset of items from a mathematics test administered to 2156 children; that is, the first three items (out of a total of 18) from the Latent GOLD demo data file "cito.dat". The column "Frequency' contains the observed frequency count for each of the eight possible answer patterns.

[INSERT TABLES 1 AND 2 ABOUT HERE]

In addition to the *J* observed dependent variables, a LC model contains a discrete latent variable. We will denote a subject's unobserved score on this latent variable by v_i, the number of LCs by *C*, and a particular class by *c*, where $c = 1, 2, \ldots, C$. The aim of a LC analysis of the data set in Table 1 could be to classify pupils into two groups, masters and non-masters, which differ with respect to the

probability of answering the test items correctly. The results obtained with a 2-class model will be used to illustrate the various components of a LC model.

LC analysis defines a model for $f(\mathbf{y}_i)$, the probability density of the multivariate response vector $\mathbf{y}_i$. In the above example this is the probability of answering the items according to one of the eight possible response patterns, for example, of answering the first two items correctly and the last one incorrectly $(y_{i1} = 1, y_{i2} = 1, y_{i3} = 0)$, which as can be seen in Table 1 equals 0.161 for the estimated 2-class model. The assumption underlying any type of LC or mixture model is that the density $f(\mathbf{y}_i)$ is a weighted average (or mixture) of the *C* class-specific densities $f(\mathbf{y}_i \mid \nu_i = c)$. This is expressed mathematically as follows:

$$f(\mathbf{y}_i) = \sum_{c=1}^{C} P(\nu_i = c) f(\mathbf{y}_i \mid \nu_i = c) \qquad (1)$$

Here, $P(\nu_i = c)$ denotes the probability that a subject belongs to LC *c*. For our small empirical example, the estimates of these (prior) class membership probabilities are .601 and .399 for class 1 and 2, respectively (see Table 2). The assumed mechanism by equation (1) is that each individual belongs to one of *C* exhaustive and mutually exclusive classes with probability $P(\nu_i = c)$ and that given membership of LC *c* an individual provides responses according to the probability density associated to this class. Table 1 shows the estimated values of $f(\mathbf{y}_i \mid \nu_i = 1)$ and $f(\mathbf{y}_i \mid \nu_i = 2)$ for our data sets. As can be seen, LC 1 has higher

probabilities for the response patterns with 2 or 3 correctly answered items, whereas LC 2 has higher probabilities for the response patterns with 0 or 1 item correct.

The classical LC model combines the assumption of equation (1) shared by all mixture models with the assumption of local independence. Local independence means that the *J* responses are mutually independent given a subject's class membership. It can be expressed as follows:

$$f(\mathbf{y}_i \mid \nu_i = c) = \prod_{j=1}^{J} f(y_{ij} \mid \nu_i = c). \qquad (2)$$

Independence implies that the joint density $f(\mathbf{y}_i \mid \nu_i = c)$ is obtained as a product of the *J* item-specific densities $f(y_{ij} \mid \nu_i = c)$. In our example, $f(y_{ij} = 1 \mid \nu_i = c)$, is the class-specific probability of giving a correct answer to item *j*. As reported in Table 2, for a subject belonging to the first LC, these equal .844, .912, and .730 for items 1, 2 and 3, respectively. The local independence assumption implies, for example, that the probability of answering the first two items correctly and the last one incorrectly for someone in LC 1 equals $.844 \times .912 \times (1\text{-}.730) = 0.208$. Note that the local independence assumption is also used in other types of latent variables models, such as in factor analysis and IRT modeling, and is thus not specific for LC analysis.

Combining the two basic equations (1) and (2) yields the following model for $f(\mathbf{y}_i)$:

$$f(\mathbf{y}_i) = \sum_{c=1}^{C} P(\nu_i = c) \prod_{j=1}^{J} f(y_{ij} \mid \nu_i = c) \ . \qquad (3)$$

To complete the model specification, we need to define the form of the conditional densities $f(y_{ij} \mid \nu_i = c)$. In the classical LC model for categorical items these are multinomial probability densities; that is,

$$f(y_{ij} \mid \nu_i = c) = \prod_{r=0}^{R_j - 1} \pi_{jrc}^{\ y^*_{ijr}} \ ,$$

where R_j is the number of categories of item j, $0 \le y_{ij} \le R_j - 1$, and $y^*_{ijr} = 1$ if $y_{ij} = r$ and 0 otherwise. Note this is a slightly complicated, but mathematically elegant, way to express that someone in LC c has a probability equal to $\pi_{jrc} = P(y_{ij} = r \mid \nu_i = c)$ of giving response r to item j. In the special case of a dichotomous response, the multinomial distribution reduces to the Bernoulli distribution with success probability $\pi_{jc} = \pi_{j1c} = P(y_{ij} = 1 \mid \nu_i = c)$. Table 2 presents these probabilities for our small example.

It is important to note that LC models can not only be used with categorical responses, but also with continuous responses and counts. The density $f(y_{ij} \mid \nu_i = c)$ could be a binomial, Poisson, or negative binomial distribution for counts, and a normal or gamma distribution for continuous responses. The mixture model for continuous response variables is sometimes referred to as the

latent profile model. The parameters of this model are the class proportions and class-specific item means and variances (μ_{jc} and σ^2_{jc}).

By comparing the *J* sets of item parameters across classes, one can name the classes. The parameter estimates presented in Table 2 show that the first class can be named the masters because pupils belonging to that class have much higher probabilities of answering the test items correctly than pupils belonging to the second non-masters class.

Similar to cluster analysis, one of the purposes of LC analysis might be to assign individuals to LCs. The probability of belonging to LC *c* given responses $\mathbf{y}_i$ – often referred to as posterior membership probability – can be obtained by the Bayes rule:

$$P(\nu_i = c \mid \mathbf{y}_i) = \frac{P(\nu_i = c) f(\mathbf{y}_i \mid \nu_i = c)}{f(\mathbf{y}_i)}. \qquad (4)$$

Table 1 reports $P(\nu_i = c \mid \mathbf{y}_i)$ for each answer pattern. For example, $P(\nu_i = 1 \mid \mathbf{y}_i)$ equals 0.774 for the (1,1,0) pattern, which is obtained as $0.601 \cdot 0.208 / 0.161$. The most common classification rule is modal assignment, which amounts to assigning each individual to the LC with the highest $f(\nu_i = c \mid \mathbf{y}_i)$. The last column of Table 1 reporting the modal assignments shows that pupils with at least 2 correct answers are assigned to class 1 and the others to class 2.

In the introduction we stated that mixture models are statistical models in which parameters are assumed to differ across LCs. But what is the statistical model used in the simple LC models discussed so far? It is the independence model: we assume responses to be independent, with different parameter values for each class. Depending on the scale type of the response variables, these parameters are Bernoulli probabilities, multinomial probabilities, normal means and variances, Poisson rates, etc.

Generalized linear models for item probabilities/means

Haberman (1979) showed that the LC model for categorical response variables can also be specified as a log-linear model for an expanded table, including the latent variable ν_i as an additional dimension. Using such a log-linear specification is equivalent to parameterizing the response probability for item *j* as follows:

$$\log\left(\frac{\pi_{jrc}}{\pi_{j0c}}\right) = \log\left(\frac{P(y_{ij}=r \mid \nu_i = c)}{P(y_{ij}=0 \mid \nu_i = c)}\right) = \alpha_{jr} + \beta_{jcr}, \qquad (5)$$

for $1 \le r \le R_j - 1$; that is, as a multinomial logistic regression model with intercepts α_{jr} and slopes β_{jcr} (note that we use the first item category, *r*=0, as baseline). One identification constraint needs to be imposed, for example,

$\beta_{j1r} = 0$ (the parameters for class 1 are fixed to 0) or $\alpha_{jr} = 0$ (intercepts are fixed to 0).

For dichotomous responses and binomial counts, the regression model could be a binary logit or probit model, for Poisson counts a log-linear model, and for continuous responses a standard linear model. These are generalized linear models (GLMs) of the form

$$g\left[E(y_{ij} \mid \nu_i = c)\right] = \alpha_j + \beta_{jc}, \qquad (6)$$

where $g[\cdot]$ is the link function transforming the expected value of y_{ij} to the linear term. For ordinal polytomous variables, one may use an ordinal regression model, such as an adjacent-category or cumulative logit model. These are models that restrict the item response probabilities π_{jrc}.

Some restricted models for categorical items

Many interesting types of restricted LC models for categorical items have been proposed which involve imposing (linear) constraints on either the conditional probabilities π_{jrc} or the logit coefficients of equation (5). One of these are probabilistic Guttman scaling models for dichotomous responses, which are LC models with $C=J+1$ classes, one for each possible total score. The idea is that apart from measurement error, class c should provide a positive (correct) answer to the c-1 easiest items and a negative (incorrect) answer to the remaining J-(c-1)

items. The various types of probabilistic Guttman models differ in constraints they impose on the measurement error. The simplest and most restricted model is the Proctor (1970) model. Table 3 presents the parameter estimates obtained when fitting the Proctor model to the data set in Table 1. As can be seen, the probability of a correct response is either 0.833 or 0.167=1-0.833. The measurement error – or the probability of giving a response which is not in agreement with the class – is estimated to be equal to 0.167. Whereas the Proctor model assumes that the measurement error is constant across items and classes, less restricted models can be defined which allow the error probabilities to differ across items, classes, or both (see, e.g., Dayton, 1999). Note that these equality constraints on the error probabilities can also be defined using linear constraints on the logit parameters: $\alpha_{j1} = 0$ and $\beta_{j1c} = -\beta^*$ for $c \leq j$ and $\beta_{jc} = \beta^*$ otherwise.

[INSERT TABLE 3 ABOUT HERE]

Croon (1990) proposed a restricted LC model that similar to non-parametric IRT (Sijtsma and Molenaar, 2000) assumes monotonic item response functions; that is, $\pi_{j1c} \leq \pi_{j1,c+1}$, or, equivalently, $\beta_{j1c} \leq \beta_{j1,c+1}$. A more restricted version, in which not only classes but also items are ordered, is obtained by imposing the additional set of restriction $\pi_{j+11c} \leq \pi_{j1c}$; that is, by assuming double monotony. Vermunt (2001) discussed various generalizations of these models.

Various authors described the connection between restricted LC analysis and parametric IRT modeling (see, for example, Heinen, 1996; Lindsay, Clogg, and Grego, 1991); that is, IRT models with a discrete specification of the distribution of the underlying trait or ability can be defined as LC models with restrictions on the logistic parameters. The key restriction is $\beta_{jrc} = \beta^*_{jr} \cdot \theta_c$ for nominal items and $\beta_{jrc} = \beta^*_j \cdot r \cdot \theta_c$ for ordinal items, where the θ_c are LC locations representing the *C* possible values of the discretized latent trait. These locations may be fixed a priori, for example, at -2, -1, 0, 1, and 2 in the case of *C*=5, but may also be treated as free parameters to be estimated. Depending on whether the items are dichotomous, ordinal, or nominal, this yields a 2-parameter logistic, generalized partial credit, or nominal response model. Further restrictions involve equating β^*_j across items, yielding Rasch and partial credit models, and imposing across category and across item restrictions on α_{jr} parameters as in rating scale models for ordinal items.

Models with explanatory variables

The most important extension of the LC models discussed so far is the possibility to include explanatory variables (covariates) affecting the responses (Wedel and DeSarbo, 1994) or the class memberships (Dayton and Macready, 1988).

Denoting the vector with explanatory variables for subject *i* by $\mathbf{x}_i$, the LC model of interest can be formulated as follows:

$$f(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{c=1}^{C} P(\nu_i = c \mid \mathbf{x}_i) \prod_{j=1}^{J_i} f(y_{ij} \mid \nu_i = c, \mathbf{x}_{ij}) . \qquad (7)$$

The main difference compared to the model defined in equation (3) is that now we have a model for $f(\mathbf{y}_i \mid \mathbf{x}_i)$ − the conditional density of $\mathbf{y}_i$ given $\mathbf{x}_i$.

Similar to the regression models for the response variables introduced in equations (5) and (6), one can define a mixture regression model with explanatory variables; i.e.,

$$g\left[E(y_{ij} \mid \nu_i = c, \mathbf{x}_{ij})\right] = \alpha_c + \sum_{p=1}^{P} \beta_{pc}\, x_{ijp} . \qquad (8)$$

As before, y_{ij} may refer to the response on item *j* by pupil *i*, in which case the explanatory variables in a long (two-level) format data file will consist of a design matrix defining the item parameters. For our small example with 3 items, the data file will contain 3 records per person--one per item. Each record has 4 columns with the first column containing an identifier variable linking the records, the second column indicating the response to a single item, and 2 additional columns for dummy-coded predictors to indicate which item the response is for. Taking the third item as the reference item, the first predictor takes on the value 1 for a record containing the first item response and 0 for the other records, and the

second predictor equals 1 for the second response and 0 for the other two responses.

However, the model in equation (8) can also be used for many other purposes. In fact, it is a model for analyzing two-level data sets, where regression parameters are allowed to differ across LCs (of higher-level units). For example, y_{ij} could be the test score of pupil *j* belonging to school *i*, and $\mathbf{x}_{ij}$ a set of pupil characteristics (e.g., IQ). A mixture regression model would identify LCs of schools with different intercepts and different effects of child characteristics on the test scores. Another possible application is in the analysis of longitudinal data, where *j* is a time point for subject *i*, and where vector $\mathbf{x}_{ij}$ contains time variables. This yields a LC growth model in which subjects are grouped based on their developmental trajectories (Vermunt, 2007). A fourth possible application is in experiments in which subjects are observed in multiple conditions, such as in conjoint studies. The mixture regression model can be used to group subjects based on their reactions on the experimental conditions. In fact, in each of these application types, the LC model is used as a random-coefficient model without parametric assumptions about the distribution of the random effects (Aitkin, 1999; Vermunt and Van Dijk, 2001).

As shown in equation (7), an individual's class membership may also be predicted using covariates. This is achieved by defining a multinomial logistic regression model for $P(\nu_i = c \mid \mathbf{x}_i)$:

$$\log \frac{P(\nu_i = c \mid \mathbf{x}_i)}{P(\nu_i = 1 \mid \mathbf{x}_i)} = \gamma_{0r} + \sum_{q=1}^{Q} \gamma_{pc} x_{iq} \ .$$

Strongly related are multiple-group LC models (Hagenaars and McCutcheon, 2002; Magidson and Vermunt, 2004; Kankaraš, Moors, Vermunt, 2010). These can be defined by using the grouping variable as a nominal explanatory in the model.

Three-step latent class analysis

Rather than including covariates directly within the estimated LC model, one may also use the following type of three-step approach:

1. Perform model selection and estimation in the usual way, thus without the inclusion of covariates;
2. Obtain class assignments *w* using the selected model from step 1, as explained in equation (4);
3. Perform subsequent analyses with covariates or other types of external variables using the class assignments *w* of step 2.

Although this stepwise approach of separating the LC analysis from the analyses one would like to do after the latent classes are constructed is very practical and intuitive, it is also problematic. More specifically, as a result of the classification errors introduced in step 2, it yields underestimated associations between external

variables and latent classes. The larger the classification errors, the larger the bias in the estimates of these associations.

However, building upon the work by Bolck, Croon, and Hagenaars (BCH, 2004), Vermunt (2010a) proposed a solution to this problem. He showed how to perform a valid step-3 analysis by adjusting for the classification errors introduced in step 2. Basically, a new LC model is estimated in which the class assignments *w* are used as the single indicator with known conditional response probabilities $P(w|\nu_i = c)$. This adjusted step-3 analysis can not only be used with covariates predicting class membership (via a logistic model), but also for investigating how classes differ with respect to a distal outcome variable (Bakk, Tekle, and Vermunt, 2013). Bakk and Vermunt (2016) recommended using the more robust BCH adjustment method for continuous distal outcomes (dependent variables), while for covariates and categorical distal outcomes the maximum likelihood (ML) approach is preferred.

Extensions

The most common model-fitting strategy in LC analysis is to increase the number of classes until the local independence assumption holds. This may, however, yield solutions which are difficult to interpret. One alternative approach is to relax the local independence assumption by allowing for associations between particular item pairs. Hagenaars (1988) showed how to define LC models with local dependencies for categorical responses. With continuous responses this is easily achieved by using multivariate instead of univariate normal distributions for locally dependent items (see, e.g., McLachlan and Peel, 2000, and Vermunt and Magidson, 2002).

Another alternative strategy involves increasing the number of discrete latent variables instead of the number of LCs, which is especially useful if the items measure several dimensions, such as different math subskills. This so-called discrete factor modeling approach (Magidson & Vermunt, 2001) is a special case of the path modeling approach for discrete latent variables developed by Hagenaars (1990) and Vermunt (1997). Many other interesting models can be defined within this framework, such as latent Markov (or transition) models for the analysis of longitudinal data (Collins and Lanza, 2010; Van de Pol and Langeheine, 1990; Vermunt, Tran, and Magidson, 2008) and LC models for cognitive diagnosis (De la Torre and Douglas 2004).

Another interesting extension is the LC tree approach (van den Bergh, van Kollenburg, Vermunt, 2018), in which after starting with a small number of classes at the root of the tree, a hierarchical structure of mutually linked classes is obtained by sequentially splitting classes into two subclasses as long as the model fit improves. This approach, which is similar to (divisive) hierarchical clustering, can also be combined with mixture growth modeling and three-step LC analysis.

Various types of models have been developed that contain both discrete and continuous latent variables, examples of which include mixture factor models (Yung, 1997; McLachlan and Peel, 2000), mixture structural equation models (Dolan and Van der Maas, 1997), and mixture IRT models (Rost, 1990). Recently, extensions of mixture factor modeling have been proposed for detecting

sources of measurement non-invariance in intensive longitudinal data, multiple-group data for many groups, and multilevel data (De Roover et al. 2017; De Roover, Vermunt, and Ceulemans, 2022; Vogelsmeier et al., 2019).

Another important extension is the multilevel LC model (Vermunt, 2003, 2010b). One of its variants is a model with discrete latent variables at multiple levels of a hierarchical structure: e.g., children belong to LCs with different performances on a set of test items, and schools belong to LCs with different distributions of children across the child-level performance classes. Multilevel LC models can be used for the analysis of two-level multivariate and three-level univariate response data.

Lanza, Coffman, and Xu (2013) and Clouth et al. (2022) showed how to combine LC analysis with tools for causal inference in observational (non-experimental) studies. More specifically, Lanza et al. (2013) illustrated the use of inverse propensity weighting and matching based on propensity scores to determine the effect of a treatment on class membership. As an alternative, Clouth et al. (2022) proposed using a three-step LC approach with inverse propensity weighting.

Maximum likelihood estimation

The parameters of LC models are typically estimated by means of maximum likelihood (ML). The log-likelihood function that is maximized is based on the probability densities defined in equations (1), (2), and (3); that is,

$$\ln L = \sum_{i=1}^{N} \ln f(\mathbf{y}_i) .$$

With categorical responses one will typically group the data and construct a frequency table as we did in Table 1. The log-likelihood function for grouped data equals

$$\ln L = \sum_{k=1}^{K} n_k \ln f(\mathbf{y}_k) ,$$

where *k* is a data pattern, *K* the number of different data patterns. and n_k the cell count corresponding to data pattern *k*. Notice that only nonzero observed cell entries contribute to the log-likelihood function, a feature that is exploited by several more efficient LC software packages that have been developed within the past few years (see the "Software" section for more discussion of these packages).

One of the problems in the estimation of LC models for discrete y_{ij} is that model parameters may be nonidentified, even if the number of degrees of freedom – the number of independent cells in the *J*-way cross-tabulation minus the number of free parameters – is larger or equal to zero. Nonidentification means that different sets of parameter values yield the same maximum of the log-likelihood

function or, worded differently, that there is no unique set of parameter estimates. The formal identification check is via the Jacobian matrix (matrix of first derivatives of $f(\mathbf{y}_i)$), which should be column full rank. Another option is to estimate the model of interest with different sets of starting values. Except for local solutions (see below), an identified model gives the same final estimates for each set of the starting values.

Although there are no general rules with respect to the identification of LC models, it is possible to provide certain minimal requirements and point to possible pitfalls. For an unrestricted LC analysis, one needs at least three responses (y_{ij}'s) per individual, but if these are dichotomous, no more than two LCs can be identified. One has to be careful with four dichotomous response variables, in which case the unrestricted three-class model is not identified, even though it has a positive number of degrees of freedom. With five dichotomous items, however, even a five-class model is identified. Usually, it is possible to achieve identification by constraining certain model parameters.

A second problem associated with the estimation of LC models is the presence of local maxima. The log-likelihood function of a LC model is not always concave, which means that hill-climbing algorithms may converge to a different maximum depending on the starting values. Usually, we are looking for the global maximum. The best way to proceed is, therefore, to estimate the model with different sets of random starting values. Typically, several sets converge to

the same highest log-likelihood value, which can then be assumed to be the ML solution. Current LC analysis software packages have automated the use of multiple sets of random starting values to reduce the probability of getting a local solution.

Another problem in LC modeling is the occurrence of boundary solutions, which are probabilities equal to 0 (or 1) or logit parameters equal to minus (or plus) infinity. These may cause numerical problems in the estimation algorithms, occurrence of local solutions, and complications in the computation of standard errors and number of degrees of freedom of the goodness-of-fit tests. Boundary solutions can be prevented by imposing constraints or by taking into account other kinds of prior information on the model parameters.

The most popular methods for solving the ML estimation problem are the expectation-maximization (EM) and Newton-Raphson (NR) algorithms. EM is a very stable iterative method for ML estimation with incomplete data. NR is a faster procedure that, however, needs good starting values to converge. The latter method makes use of the matrix of second-order derivatives of the log-likelihood function, which is also needed for obtaining standard errors of the model parameters.

Model selection issues

The goodness-of-fit of LC models for categorical responses can be tested using Pearson and likelihood-ratio chi-squared tests. The latter is defined as

$$L^2 = 2\sum_{k=1}^{K} n_k \ln \frac{n_k}{N \cdot f(\mathbf{y}_k)}.$$

As in log-linear analysis, the number of degrees of freedom (*df*) equals the number of cells in the frequency table minus 1, minus the number of independent parameters. In an unrestricted LC model,

$$df = \prod_{j=1}^{J} R_j - C \cdot \left[1 + \sum_{j=1}^{J} (R_j - 1) \right].$$

Although it is no problem to estimate LC models with 10, 20, or 50 indicators, in such cases, the frequency table may become very sparse and, as a result, asymptotic *p*-values can no longer be trusted. An elegant but somewhat time-consuming solution to this problem is to estimate the *p*-values by parametric bootstrapping. Another option is to assess model fit in lower-order marginal tables (e.g., in the two-way marginal tables). Magidson and Vermunt (2004) refer to Pearson chi-squared statistics in two-way tables as bivariate residuals (or BVRs).

Even though models with C and $C + 1$ are nested, one cannot test them against each other using a standard likelihood-ratio ($-2 \ln L$ difference) test because it does not have an asymptotic chi-squared distribution. A solution to this problem is to approximate its sampling distribution using bootstrapping. But since

this method is computationally demanding, usually alternative methods are required for comparing models with different numbers of classes. One popular method is the use of information criteria, such as the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC), and the AIC3 (a variant of the AIC which uses a penalty of 3 instead of 2 per parameter), where smaller values indicate better model fit.

Usually, we are not only interested in goodness-of-fit but also in the performance of the modal classification rule (see equation 4). The estimated proportion of classification errors under modal classification equals

$$E = \sum_{i=1}^{N} \frac{1}{N} \left\{ 1 - \max\left[P(\nu_i = c \mid \mathbf{y}_i) \right] \right\}.$$

This number can be compared to the proportion of classification errors based on the unconditional probabilities $P(\nu_i = c)$, yielding a reduction of errors measure

$$\lambda = 1 - \frac{E}{1 - \max\left[P(\nu_i = c) \right]}.$$

The closer this nominal R^2-type measure is to 1, the better the classification performance of a model. Other types of classification error reduction measures have been proposed based on entropy or qualitative variance (Collins and Lanza, 2010; Vermunt and Magidson, 2016).

[INSERT TABLE 4 ABOUT HERE]

To illustrate the use of model selection statistics for deciding about the number of classes, Table 4 presents the results for 1- to 6-class models estimated with the "cito.dat" data set using all 18 items, where p-values for the goodness-of-test and the likelihood-ratio test were obtained using bootstrapping (with 500 replications). The simplest model with a non-significant L^2 goodness-of-fit value and moreover with the lowest BIC is the 3-class model, and should thus be selected according to these two statistics. The 4-class model would be selected based on the AIC3, the bootstrap likelihood-ratio (-2 ln L difference) test, and the BVR: the AIC3 is lowest, the likelihood-ratio test shows it is significantly better than the 3-class model and non-significantly worse than the 5-class model, and the largest BVR drops quite a bit when going from 3 to 4 classes, but not anymore afterwards. The AIC indicates that at least 6 classes are needed, but simulation studies have shown that it tends to overestimate the number of classes.

As also happens in this example application, different criteria typically point at different "best" models, meaning that either preference for parsimony and/or content knowledge should determine the final decision, in this example, on whether to retain the 3- or 4-class model. The last column of Table 4 reports the entropy-based R^2 value indicating how well one can predict the individuals' class memberships based on their observed responses. As in our example, this measure typically decreases with the number of classes. It should be noted that it should not be used for model selection, but only as a measure indicating how well the

selected model performs in terms of classification (similar to a reliability coefficient of a scale).

Software

One of the first LC analysis programs, MLLSA, made available by Clifford Clogg in 1977, was limited to a relatively small number of nominal variables. Today's programs can handle many more variables, as well as other scale types. For example, the LEM program (Vermunt, 1997) provides a command language that can be used to specify a large variety of models for categorical data, including LC models. Mplus is a command language–based structural equation modeling package that implements many types of LC and mixture models. In addition, routines and packages for the estimation of specific types of LC models are available for SAS, R, and Stata (see, for example, Lanza et al., 2007; Linzer and Lewis, 2011; and Skrondal and Rabe-Hesketh, 2004). Haughton, Legrand, and Woolford (2009) reviewed the Latent GOLD program and the R packages poLCA and MCLUST.

Latent GOLD (Vermunt and Magidson, 2016, 2021) is a stand-alone program that was especially developed for LC analysis, and which contains both an SPSS-like point and click user interface and a syntax language. It implements all important types of LC models, such as models for response variables of different scale types, restricted LC models, models with predictors, models with

local dependencies, models with multiple discrete latent variables, path models with discrete latent variables, latent Markov/transition models, mixture regression and mixture growth models, mixture factor analysis and IRT, multilevel LC models, three-step LC modeling, and LC tree models, as well as features for dealing with partially missing data, performing bootstrapping, generating multiple imputed data sets, performing simulation studies, and providing correct inference with complex sampling designs.

Bibliography

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55, 218–234.

Bakk, Zs., Tekle, F.B., and Vermunt, J.K. (2013). Estimating the association between latent class membership and external variables using bias adjusted three-step approaches. *Sociological Methodology*, 43, 272-311.

Bakk, Zs., and Vermunt, J.K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling*, 23, 20-31.

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

Clouth, F.J., Pauws, S., Mols, F., and Vermunt, J.K. (2022). A new three-step method for using inverse propensity weighting with latent class analysis. *Advances in Classification and Data Analysis*, in press.

Croon, M. A. (1990). Latent class analysis with ordered latent classes. British *Journal of Mathematical and Statistical Psychology*, 43, 171-192.

Dayton, C. M. (1999). *Latent class scaling analysis*. Thousand Oaks: Sage Publications.

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

De la Torre J, and Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.

De Roover, K., Vermunt, J. K., and Ceulemans, E. (2022). Mixture multigroup factor analysis for unraveling factor loading non-invariance across many groups, *Psychological Methods*, in press.

De Roover, K., Vermunt, J. K., Timmerman, M., and Ceulemans, E. (2017). Mixture simultaneous factor analysis for capturing differences in latent variables between higher-level units of multilevel data, *Structural Equation Modeling*, 24, 506-523.

Dolan, C. V., and Van der Maas, H. L. J. (1997). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, 63, 227-253.

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

Haberman, S. J. (1979). *Analysis of qualitative data: Vol. 2. New developments*. New York: Academic Press.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods and Research*, 16, 379-405.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Loglinear analysis of panel, trend and cohort data*. Newbury Park, CA: Sage.

Haughton, D., Legrand, P., and Woolford, S. (2009). Review of three latent class cluster analysis packages: Latent Gold, poLCA, and MCLUST. *The American Statistician*, 63, 81-91

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.

Kankaraš, M., Moors, G., and Vermunt, J. K. (2010). Testing for measurement invariance with latent class analysis. E. Davidov, P. Schmidt, and J. Billiet (eds), *Cross-cultural analysis: methods and applications* (pp. 359–384) Routledge, New York.

Lanza, S. T., Coffman, D. L., and Xu, S. (2013) Causal inference in latent class analysis. *Structural Equation Modeling*, 20, 361–383

Lanza, S. T., Collins, L. M., Lemmon, D. R., and Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling*, 14, 671 – 694.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & the interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer, et al. (Eds.), *Measurement and prediction* (pp. 362–472). Princeton, NJ: Princeton University Press.

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86, 96-107.

Linzer, D. A., and Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(1), 1-29.

Magidson, J., and Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology*, 31, 223–264.

McLachlan, G.J., and Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons.

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. Psychometrika, 35, 73-78.

Rost, J. (1990). Rasch models in latent classes. An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.

Sijtsma, K., and Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: Sage.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. London: Chapman & Hall/CRC.

Van den Bergh, M., van Kollenburg, G. H., and Vermunt, J. K. (2018). Deciding on the starting number of classes of a latent class tree, *Sociological Methodology*, 48, 303-336.

Van de Pol, F., and Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, 20, 213-247.

Vermunt, J. K. (1997). *Log-linear models for event histories*. Thousand Oaks, CA: Sage.

Vermunt, J. K. (2001). The use restricted latent class models for defining and testing nonparametric and parametric IRT models. *Applied Psychological Measurement*, 25, 283-294.

Vermunt, J. K.(2003). Multilevel latent class models. *Sociological Methodology*, 33, 213-239.

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

Vermunt, J. K. (2010a). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18, 450-469.

Vermunt, J.K., and Magidson, J. (2002) Latent class cluster analysis. J. Hagenaars and A. McCutcheon (eds.). *Applied latent class analysis* (pp. 89-106). Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2013). *Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2021). *Upgrade manual for Latent GOLD Basic, Advanced/Syntax and Choice version 6.0*. Arlington, MA: Statistical Innovations Inc.

Vermunt, J. K., and Van Dijk. L. (2001). A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter*, 13, 6-13.

Vogelsmeier, L. V. D. E., Vermunt, J. K., Böing-Messing, F., and De Roover, K. (2019). Continuous-time latent Markov factor analysis for exploring measurement model changes across time, *Methodology*, 15(Supplement), 29-42.

Wolfe, J. H. (1970). Pattern clustering by mulltivariate cluster analysis. *Multivariate Behavioral Research*, 5, 329-350.

Wedel, M., and DeSarbo, W. (1994). A review of recent developments in latent class regression models. R.P. Bagozzi (ed.), *Advanced methods of marketing research* (pp. 352-388). Cambridge, MA: Blackwell Publishers.

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, 62, 297-330.

Further reading

Reading Material	Description
Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley.	This is a text book on latent class analysis and latent transition models. It deals with models for categorical responses.
Hagenaars, J. A., and McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.	This edited volume contains contributions of the main researchers involved in the development of LC models before 2000. It can be seen as the state-of-art of the field around 2000.
Magidson, J., and Vermunt, J. K. (2004). Latent class models. D. Kaplan (ed.), *The Sage handbook of quantitative methodology for the social sciences*, (pp. 175-198). Thousand Oakes: Sage Publications.	This handbook chapter gives more detailed explanation on the practical application of latent class models. An updated version from 2016 and accompanying links to video can be found at www.jeroenvermunt.nl.
Vermunt, J. K. (2010b). Mixture models for multilevel data sets. J. Hox & J. K. Roberts (eds.). *The handbook of advanced multilevel analysis* (pp. 59-81). New York: Routledge.	This handbook chapter discusses the most important types of LC models for multilevel data sets. These include mixture regression models for 2-level and 3-level data sets and LC models with discrete mixtures at multiple levels.

Vermunt, J. K., Tran, B., and Magidson, J. (2008). Latent class models in longitudinal research. In: S. Menard (ed.) *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 373-385). Burlington, MA: Elsevier.	This handbook chapter gives an overview of the various types of LC models for longitudinal data sets. These include mixture growth models and latent transition models.

Table 1: Small data set with three dichotomous responses

y_{i1}	y_{i2}	y_{i3}	Frequency	$f(\mathbf{y}_i)$	$f(\mathbf{y}_i \mid \nu_i = 1)$	$f(\mathbf{y}_i \mid \nu_i = 2)$	$f(\nu_i = 1 \mid \mathbf{y}_i)$	$f(\nu_i = 2 \mid \mathbf{y}_i)$	Modal
0	0	0	239	0.111	0.004	0.272	0.020	0.980	2
0	0	1	101	0.047	0.010	0.102	0.128	0.872	2
0	1	0	283	0.131	0.038	0.271	0.175	0.825	2
0	1	1	222	0.103	0.104	0.102	0.605	0.395	1
1	0	0	105	0.049	0.020	0.092	0.248	0.753	2
1	0	1	100	0.046	0.054	0.035	0.703	0.297	1
1	1	0	348	0.161	0.208	0.091	0.774	0.226	1
1	1	1	758	0.352	0.562	0.034	0.961	0.039	1

Note: These are the first three items (out of a total of 18) from the Latent GOLD demo data file called “cito.dat”.

Table 2: Parameters (class proportions and probability of a correct answer) obtained with a 2-class model for data in Table 1

	c=1	c=2
$P(\nu_i = c)$	0.601	0.399
$\pi_{11c} = P(y_{i1} = 1 \mid \nu_i = c)$	0.844	0.252
$\pi_{21c} = P(y_{i2} = 1 \mid \nu_i = c)$	0.912	0.499
$\pi_{31c} = P(y_{i3} = 1 \mid \nu_i = c)$	0.730	0.273

Table 3: Parameters (class proportions and probability of a correct answer) obtained with Proctor model for data in Table 1

	c=1	c=2	c=3	c=4
$P(\nu_i = c)$	0.160	0.155	0.126	0.559
$\pi_{21c} = P(y_{i2} = 1 \mid \nu_i = c)$	0.167	0.833	0.833	0.833
$\pi_{11c} = P(y_{i1} = 1 \mid \nu_i = c)$	0.167	0.167	0.833	0.833
$\pi_{31c} = P(y_{i3} = 1 \mid \nu_i = c)$	0.167	0.167	0.167	0.833

Table 4: Fit measures for the LC models estimated using all 18 items of the "cito.dat" data set.

Model	BIC	AIC	AIC3	L^2	Bootstrap p-value L^2	Largest BVR	-2 ln L Difference	Bootstrap p-value -2 ln L Difference	Entropy R^2
1 Class	44667.15	44564.98	44582.98	15115.33	0.00	213.86			
2 Classes	41055.03	40845.02	40882.02	11357.36	0.01	43.34	3757.96	0.00	0.80
3 Classes	<u>40779.08</u>	40461.22	40517.22	<u>10935.57</u>	<u>0.38</u>	11.89	421.79	0.00	0.69
4 Classes	40830.76	40405.06	<u>40480.06</u>	10841.40	0.41	<u>6.35</u>	<u>94.17</u>	<u>0.00</u>	0.61
5 Classes	40929.23	40395.69	40489.69	10794.04	0.37	5.74	47.37	0.07	0.60
6 Classes	41029.75	<u>40388.36</u>	40501.36	10748.70	0.34	5.92	45.33	0.12	0.59

Note: The selected model is underlined. BIC, AIC, and AIC3 select the model with the lowest value, L^2 the simplest model with a non-significant value, largest BVR the simplest model showing a large drop compared to the previous model, and -2 ln L difference the most complex model with a significant value. Entropy R^2 should not be used for model selection.