Applications of Latent Class Analysis in Social Science Research

Jeroen K. Vermunt

Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands

Abstract. An overview is provided of recent developments in the use of latent class (LC) models in social science research. Special attention is paid to the application of LC analysis as a factor-analytic tool and as a tool for random-effects modeling. Furthermore, an extension of the LC model to deal with nested data structures is presented.

1 Introduction

Latent class (LC) analysis was introduced by Lazarsfeld in 1950 as a way of formulating latent attitudinal variables from dichotomous survey items (see [11]). During the 1970s, LC methodology was formalized and extended to nominal variables by Goodman [6] who also developed the maximum likelihood algorithm that has served as the basis for most LC programs. It has, however, taken many years till the method became a generally accepted tool for statistical analysis. The history and state-of-art of LC analysis in social science research is described in the recent volume "Applied Latent Class Analysis" edited by Hagenaars and McCutcheon [8].

Traditionally, LC models were used as clustering and scaling tools for dichotomous indicators. Scaling models, such as the probabilistic Guttman scales, involved specification of simple equality constraints on the item conditional probabilities in order to guarantee that the latent variable would capture a single underlying dimension. A more recent development is to parametrize the item conditional by means of logit models, yielding restricted variants of LC analysis which are similar to latent trait models (see [4], [9], and [19]). The log-linear modeling framework with latent variables implemented in the LEM software package yields are general class of probability models (graphical models) for categorical observed and latent variables in which each of the model probabilities can be restricted by logit constraints (see [7] and [17]). The LEM framework contains most types of LC models for categorical observed variables as special cases, including models with several latent variables, models with covariates, models for ordinal variables, models with local dependencies, causal models with latent variables, and latent Markov models.

A very much related field is the field of finite mixture (FM) modelling (see[15]). Traditionally, finite mixture models dealt with continuous outcome variables. The underlying idea of LC and FM models is, however, the same: the

population consists of a number of subgroups which differ with respect to the parameters of the statistical model of interest. It is, therefore, not surprising that in recent years, the fields of LC and FM modeling have come together and that the terms LC model and FM model have become interchangeable with each other. For example, mixture model clustering and mixture regression analysis are now also known as LC clustering and LC regression analysis.

The software package Latent GOLD (see [20] and [21]) implements the most important social science application types of LC and FM models – clustering, scaling, and random-effects modeling – in three modules: LC cluster, LC factor, and LC regression. What is very important for applied researchers is that the models are implemented in a SPSS-like graphical user interface. The use of LC analysis for clustering purposes is also well-known outside the social science field. LC factor is a factor-analytic tool for discrete or mixed outcome variables (see [12]). LC regression makes it possible to take into account unobserved population heterogeneity with respect to the coefficients of a regression model (see [23] and [24]). In this paper, I will explain the basic ideas underlying the LC factor and LC regression models and present several empirical examples.

LC models are models for two-level data structures. The data consists of a set of indicators or a set of repeated responses which are nested within individuals. Recently, models have been proposed for nested data structures consisting of more than two levels, such as repeated measures nested within persons and persons nested with groups – teams, countries, or organizations (see [18]). At the end of this paper, I will pay attention to this hierarchical or multilevel extension of the LC model and present a procedure called upward-downward algorithm that can be used to solve the maximum likelihood estimation problem.

2 The LC Factor Model

Let us start introducing some notation. Let y_{ik} denote the realized value of person i on the kth indicator, item, or response variable. The total number of response variables is denoted by K. A category of the ℓth latent class variable will be denoted as x_ℓ, its total number of categories as T_ℓ, and the total number of latent variables by L.

The standard LC model that I will refer to as the LC cluster model assumes that responses are independent of each other given a single latent variable with T_1 unordered categories. The density of $\mathbf{y}_i$ is defined as

$$f(\mathbf{y}_i) = \sum_{x_1=1}^{T_1} P(x_1) \prod_{k=1}^{K} f(y_{ik}|x_1),$$

where the exact form of the class-specific densities $f(y_{ik}|x_1)$ depends on the scale type of the response variable concerned. The $f(y_{ik}|x_1)$ are taken from the exponential family.

The main difference between the LC factor and the LC cluster model is that the former may contain more than one latent variable. Another difference is

that in the factor model the categories of the latent variables are assumed to be ordered. Thus, rather than working with a single nominal latent variable, here we work with one or more dichotomous or ordered polytomous latent variables (Magidson and Vermunt in [12]). The advantage of this approach is that it guarantees that each of the factors capture no more than one dimension.

The primary difference between our LC factor and the traditional factor analysis model is that the latent variables (factors) are assumed to be dichotomous or ordinal as opposed to continuous and normally distributed. Because of the strong similarity with traditional factor analysis, this approach is called LC factor analysis. There is also a strong connection between LC factor models and item response or latent trait models. Actually, LC factor models are discretized variants of well-known latent trait models for dichotomous and polytomous items (see [9], [19], and [22]).

As in maximum likelihood factor analysis, modeling under the LC factor approach can proceed by increasing the number of factors until a good fitting model is achieved. This approach to LC modeling provides a general alternative to the traditional method of obtaining a good fitting model by increasing the number of latent classes. In particular, when working with dichotomous uncorrelated factors, there is an exact equivalence in the number of parameters of the two models. A LC factor model with 1 factor has the same number of parameters as a 2-class LC cluster model, a model with 2 factors as a 3-class model, a model with 3 factors as a 4-class model, etc. Thus, in an exploratory analysis, rather than increasing the number of classes one may instead increase the number of factors until an acceptable fit is obtained.

2.1 A Two-factor Model for Nominal Indicators

To illustrate the LC factor model, let us assume that we have a two-factor model for four nominal categorical indicators. The corresponding probability structure is of the form

$$P(y_{i1}, y_{i2}, y_{i3}, y_{i4}) = \sum_{x_1=1}^{T_1} \sum_{x_2=1}^{T_2} P(x_1, x_2) \prod_{k=1}^{4} P(y_{ik}|x_1, x_2).$$

The conditional response probabilities $P(y_{ik}|x_1, x_2)$ are restricted by means of multinomial logit models with linear terms

$$\eta(y_{ik}|x_1, x_2) = \beta_{0y_k} + \beta_{1y_k} \cdot v_{x_1} + \beta_{2y_k} \cdot v_{x_2}.$$

Because the factors are assumed to be ordinal (or discrete interval) variables, the two-variable terms are restricted by using fixed category scores for the levels of the factors. Note that the factors are treated as metric variables, which are, however, not continuous but discrete. The scores v_{x_ℓ} for the categories of the ℓth factor are equidistant scores ranging from 0 to 1. The first level of a factor gets the score 0 and the last level the score 1. The parameters describing the strength of relationships between the factors and the indicators – here, β_{1y_k} and β_{2y_k} – can be interpreted as factor loadings.

Note that the above logit model does not include the three-variable interaction term of the two factors and the indicator. These higher-order terms are excluded from the model in order to be able to distinguish the various dimensions. If we would include the three-variable interaction term, our two-factor model would be equivalent to an unrestricted 4-cluster model. By excluding this term, we obtain a restricted 4-cluster model in which each of the four clusters can be conceived as being a combination of two factors.

In the standard LC factor model, the factors are specified to be dichotomous, which means that the scoring of the factor levels does not imply a constraint. An important extension of this standard model is, however, increasing the number of levels of a factor, which makes it possible to describe more precisely the distribution of the factor concerned. Note that the levels of the factors remain ordered by the use of fixed equal-interval category scores in their relationships with the indicators. Therefore, each additional level costs only one degree of freedom; that is, there is one additional class size to be estimated.

In the default setting, the factors are assumed to be independent of one another. This is specified by the appropriate logit constraints on the latent probabilities. In the two-factor case, this involves restricting the linear term in the logit model for $P(x_1, x_2)$ by

$$\eta_{x_1 x_2} = \gamma_{x_1} + \gamma_{x_2}.$$

Working with correlated factors is comparable to performing an oblique rotation. The association between each pair of factors is described by a single uniform association parameter:

$$\eta_{x_1 x_2} = \gamma_{x_1} + \gamma_{x_2} + \gamma_{12} \cdot v_{x_1} \cdot v_{x_2}.$$

It should be noted that contrary to traditional factor analysis, the LC factor model is identified without additional constraints, such as setting certain factor loadings equal to zero. Nevertheless, it is possible to specify models in which factor loadings are fixed to zero. Together with the possibility to include factor correlation in the model, this option can be used for a confirmatory factor analysis. Other extensions are the use of indicators which are ordinal, continuous, or counts, the inclusion of local dependencies, and the inclusion of covariates affecting the factors.

Zhang [25] proposed a LC model with several latent variables called hierarchical LC model that is similar to our LC factor model presented. Three important differences are that his factors are nominal instead of ordinal, that indicators are allowed to be related to only one factor, and that factor correlations are induced by higher-order factors.

2.2 Graphical Displays

Magidson and Vermunt [12] proposed a graphical display similar to the one obtained in correspondence analysis to depict the results of a LC factor analysis.

These displays help in detecting which indicators are related to which factors. The measures that are display are derived from the posterior factor means.

Case i's posterior mean on factor ℓ equals

$$E(v_{i\ell}) = \sum_{x_\ell=1}^{T_\ell} v_{x_\ell} P(x_\ell|\mathbf{y}_i).$$

The basic idea is to aggregate these posterior means (factor scores) and plot the resulting numbers in a two-dimensional display. Note that these numbers will be in the 0-1 range because the category score v_{x_ℓ} is 0 for the lowest factor level and 1 for the highest level. The most important aggregation is within categories of the indicators; that is,

$$E(v_\ell|y_k) = \frac{\sum_{i=1}^{N} E(v_{i\ell}) I(y_{ik} = y_k)}{\sum_{i=1}^{N} I(y_{ik} = y_k)},$$

where $I(y_{ik} = y_k)$ equals 1 if person i's value on indicator k is y_k, and 0 otherwise. This yields the mean of factor ℓ for persons who give response y_k on indicator k. These category-specific factor means will be very different if an indicator is strongly related to a factor.

Aggregation can be done for any relevant subgroup and not just for categories of the indicators. Often it is useful to depict the position of groups formed on the basis of socio-demographic characteristics. It is also possible to depict the posterior means of individual cases in the plot, yielding what is sometimes referred to as a bi-plot.

2.3 Application: Types of Survey Respondents

We will now consider an example that illustrates how the LC factor model can be used with nominal variables. It is based on the analysis of 4 variables from the 1982 General Social Survey given by McCutcheon [13] to illustrate how standard LC modeling can be used to identify different types of survey respondents. Two of the variables ascertain the respondent's opinion regarding the purpose of surveys (Purpose) and how accurate they are (Accuracy), and the others are evaluations made by the interviewer of the respondent's levels of understanding of the survey questions (Understanding) and cooperation shown in answering the questions (Cooperation). McCutcheon initially assumed the existence of 2 latent classes corresponding to 'ideal' and 'less than ideal' types. The study included separate samples of white and black respondents. Here, I use the data of the white respondents only.

The two-class LC model – or, equivalently, the 1-factor LC model – does not provide a satisfactory description of this data set ($L^2 = 75.5$; $df = 22$; $p < .001$). Two options for proceeding are to increase the number of classes or to increase the number of factors. The 2-factor LC model fits very well ($L^2 = 11.1$; $df = 15$; $p = .75$), and also much better than the unrestricted 3-class model ($L^2 = 22.1$; $df = 15$; $p = .11$) that was selected as final model by McCutcheon.

Table 1. Logit Parameter Estimates for the 2-Factor LC Model as Applied to the GSS'82 Respondent-Type Items

Item	Category	x_1	x_2
Purpose	good	-1.12	2.86
	depends	0.26	-0.82
	waste	0.86	3.68
Accuracy	mostly true	-0.52	-1.32
	not true	0.52	1.32
Understanding	good	-1.61	0.58
	fair/poor	1.61	-0.58
Cooperation	interested	-2.96	-0.57
	cooperative	-0.60	-0.12
	impatient/hostile	3.56	0.69

The logit parameter estimates obtained from the 2-factor LC model are given in Table 1. These show the magnitude of the relationship between the observed variables and the two factors. As can be seen, the interviewers' evaluations of respondents and the respondents' evaluations of surveys are clearly different factors: Understanding and Cooperation are more strongly affected by the first factor and Purpose and Accuracy by the second factor.

Fig. 1 depicts the bi-plot containing the category-specific factor means of the four indicators. The plot shows even more clearly than the logit coefficients that the first dimension differentiates between the categories of Understanding and Cooperation and the second between the categories of Purpose and Accuracy.

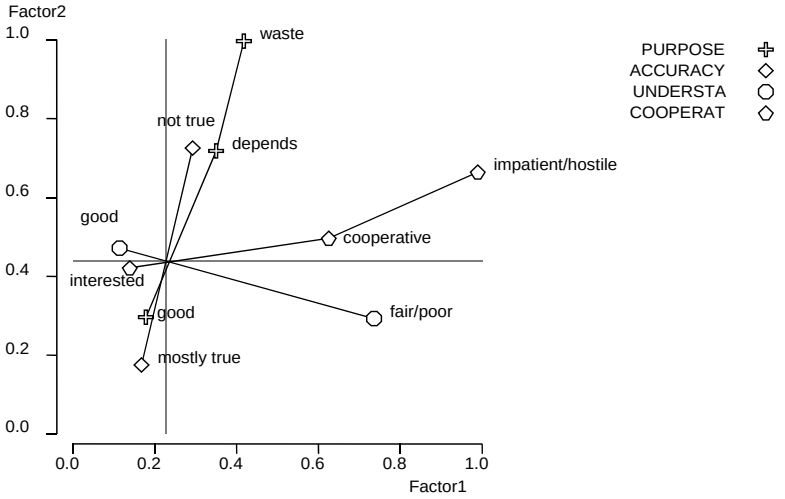
3 LC Regression Analysis

One of the differences between LC regression analysis and the other forms of LC analysis discussed so far is that it concerns a model for a single response variable. This response variable is explained by a set of predictors, where the predictor effects may take on different values for each latent class (see [10], [23], [24], and section 13.2 in [1]).

An important feature of LC regression models is that for each case we may have more than one observation. These multiple observations may be experimental replications, repeated measurements at different time points or occasions, clustered observations, or other types of dependent observations. Here, I will use the term replications, where the replication number will be denoted by k. The value of the response variable for case i at replication k is denoted by y_{ik}. The number of replications, which is not necessarily the same for all cases, is denoted by K_i. Because we are dealing with models with a single latent variable, we drop the index ℓ from x_ℓ.

Note that I am describing a two-level data structure in which a predictor may either have the same value or change its value across replications. The former

Fig. 1. Graphical Display of Category-Specific Posterior Factor Means for the 2-Factor LC Model as Applied to the GSS'82 Respondent-Type Items



are the higher-level or level-2 predictors and the latter are lower-level or level-1 predictors. Here, k indexes the (dependent) lower-level observations within a certain higher-level observation. Level-1 predictors will be denoted as z_{ikp} and level-2 predictors as w_{iq}. The LC regression model can be used to define (non-parametric) two-level or random-coefficient models. Using k as an index for time points or time intervals, one obtains models for longitudinal data, such as growth or event-history models with non-parametric random coefficients (see [17] and [23]).

Using the same notation as above, the probability structure underlying the LC regression model can be defined as

$$f(\mathbf{y}_i|\mathbf{w}_i, \mathbf{z}_i) = \sum_{x=1}^{T_1} P(x) \prod_{k=1}^{K_i} f(y_{ik}|x, \mathbf{w}_i, \mathbf{z}_{ik}) \,.$$

Similarly to other LC models, replications are assumed to be independent given class membership. For nominal or ordinal dependent variables, the probability density $f(y_{ik}|x, \mathbf{w}_i, \mathbf{z}_{ik})$ will usually be assumed to be multinomial, for continuous variables, univariate normal, and for counts, Poisson or binomial.

The linear predictor in $f(y_{ik}|x, \mathbf{w}_i, \mathbf{z}_{ik})$ equals

$$\eta(y_{ik}|x, \mathbf{w}_i, \mathbf{z}_{ik}) = \beta_{0x} + \sum_{p=1}^{P} \beta_{px}\, z_{ikp} + \sum_{q=1}^{Q} \beta_{P+q} w_{iq}$$

where P and Q denote the number of level-1 and level-2 predictors. This regression model contains a class-specific intercept, P class-specific regression coefficients, and Q class-independent regression coefficients. The $P+1$ coefficient that are class dependent are random coefficients.

The conceptual equivalence between the LC regression model and a two-level random-coefficient model becomes even clearer if one realizes that it is possible to compute the means, variances, and covariances of the class-specific coefficients from the standard LC class output. These are obtained by elementary statistics calculus:

$$\mu_p = \sum_{x=1}^{T} \beta_{px} P(x)$$

$$\tau_{pp'} = \sum_{x=1}^{T} \left(\beta_{px} - \mu_p\right) \left(\beta_{p'x} - \mu_{p'}\right) P(x),$$

This shows that LC regression analysis results can be summarized to yield information that is equivalent to the one obtained in regression models with random coefficients coming from a normal distribution; that is, it possible to obtain the mean vector and the covariance matrix of the random coefficients.

3.1 Application: Longitudinal Study on Attitudes Towards Abortion

In order to demonstrate the non-parametric random-coefficient model, I used a data set obtained from the data library of the Multilevel Models Project, at the Institute of Education, University of London. The data consist of 264 participants in 1983 to 1986 yearly waves from the British Social Attitudes Survey (see [14]). It is a three-level data set: individuals are nested within constituencies and time-points are nested within individuals. I will only make use of the latter nesting, which means that we are dealing with a standard repeated measures model. As was shown by Goldstein [5], the highest level variance – between constituencies – is so small that it can reasonably be ignored. Below, I will show how to extend the LC model to deal with higher-level data structures.

The dependent variable is the number of yes responses on seven yes/no questions as to whether it is woman's right to have an abortion under a specific circumstance. Because this variable is a count with a fixed total, it most natural to work with a logit link and binomial error function. Individual level predictors in the data set are religion, political preference, gender, age, and self-assessed social class. In accordance with the results of Goldstein, I found no significant effects of gender, age, self-assessed social class, and political preference. Therefore,

Table 2. Test results for the estimated models with the attitudes towards abortion data

Model	Log-likelihood	# parameters	BIC
No random effects			
Ia. empty model	-2309	1	4623
Ib. time linear + religion	-2215	5	4458
Ic. time dummies + religion	-2188	7	4416
Ic + Random intercept			
IIa. 2 classes	-1755	9	3560
IIb. 3 classes	-1697	11	3456
IIc. 4 classes	-1689	13	3451
IId. 5 classes	-1689	15	3461
Ic + Random intercept and slope			
IIIa. 2 classes	-1745	12	3558
IIIb. 3 classes	-1683	17	3460
IIIc. 4 classes	-1657	22	3436
IIId. 5 classes	-1645	27	3441

I did not used these predictors in the further analysis. The predictors that were used are the level-1 predictor year of measurement (1=1983; 2=1984; 3=1985; 4=1986) and the level-2 predictor religion (1=Roman Catholic, 2=Protestant; 3=Other; 4=No religion). Effect coding is used for nominal predictors.

The LC regression models were estimated by means of version 3.0 of the Latent GOLD program (see [21]), which also provides the multilevel type parameters μ and $\sqrt{\tau^2}$. I started with three models without random effects: an intercept-only model (Ia), a model with a linear effect of year (Ib), and a model with year dummies (Ic). Models Ib and Ic also contained the nominal level-2 predictor religion. The test results reported in the first part of Table 2 show that year and religion have significant effects on the outcome variable and that it is better to treat year as non-linear. I proceeded by adding a random intercept. The test results show that the model with 4 classes is the best one in terms of BIC value. Subsequently, I allowed the time effect to be class specific. Again, the 4-class model turned out to be the best according to the BIC criterion.

Table 3 reports the parameter estimates for Model IIIc. The means indicate that the attitudes are most positive at the last time point and most negative at the second time point. Furthermore, the effects of religion show that people without religion are most in favor and Roman Catholics and Others are most against abortion. Protestants have a position that is close to the no-religion group.

As can be seen, the 4 latent classes have very different intercepts and time patterns. The largest class 1 is most against abortion and class 3 is most in favor of abortion. Both latent classes are very stable over time. The overall level of latent class 2 is somewhat higher than of class 1, and it shows somewhat more change of the attitude over time. People belonging to latent class 4 are very

Table 3. Parameters estimates obtained with Model IIIc for the attitudes towards abortion data

Parameter	Class 1	Class 2	Class 3	Class 4	Mean	Std.Dev.
Class size	0.30	0.28	0.24	0.19		
Intercept	-0.34	0.60	3.33	1.59	1.16	1.38
Time						
1983	0.14	0.26	0.47	-0.58	0.12	0.35
1984	-0.11	-0.46	-0.35	-1.11	-0.45	0.34
1985	-0.04	-0.44	-0.26	1.43	-0.10	0.66
1986	-0.06	0.64	0.14	0.26	0.24	0.27
Religion						
Catholic	-0.53	-0.53	-0.53	-0.53	-0.53	0.00
Protestant	0.20	0.20	0.20	0.20	0.20	0.00
Other	-0.10	-0.10	-0.10	-0.10	-0.10	0.00
No Religion	0.42	0.42	0.42	0.42	0.42	0.00

instable: at the first two time points they are similar to class 2, at the third time point to class 4, and at the last time point again to class 2 (this can be seen by combining the intercepts with the time effects). Class 4 could therefore be labelled as random responders. It is interesting to note that in a three-class solution the random-responder class and class two are combined. Thus, by going from a three- to a four-class solution one identifies the interesting group with less stable attitudes.

3.2 Application: Choice-Based Conjoint Study

The LC regression model is a popular tool for the analysis of data from conjoint experiments in which individuals rate separate product or choose between sets of products having different attributes (see [10]). The objective is to determine the effect of product characteristics on the rating or the choice probabilities or, more technically, to estimated the utilities of product attributes. LC analysis is used to identify market segments for which these utilities differ. The class-specific utilities can be used to estimate the market share of possible new products; that is, to simulate future markets.

For illustration of LC analysis of data obtained from choice-based conjoint experiments, I will use a generated data set. The products are 12 pairs of shoes that differ on 3 attributes: Fashion (0=traditional, 1= modern), Quality (0=low, 1=high), and Price (ranging from 1 to 5). Eights choice sets offer 3 of the 12 possible alternative products to 400 individuals. Each choice task consists of indicating which of the three alternatives they would purchase, with the response "none of the above" allowed as a fourth choice option.

The regression model that is used is a multinomial logit model with choice-specific predictors, also referred to as the conditional logit model. The BIC values indicated that the three-class model is the model that should be preferred. The parameter estimates obtained with the 3-class model are reported in Table 4.

Table 4. Estimates of the parameters for 3-class choice model

Parameter	Class 1	Class 2	Class 3	Mean	Std.Dev.
Class size	0.51	0.26	0.24		
Fashion	3.03	-0.17	1.20	1.77	1.37
Quality	-0.09	2.72	1.12	0.92	1.16
Price	-0.39	-0.36	-0.56	-0.42	0.08
None	1.29	0.19	-0.43	0.60	0.73

As can be seen, Fashion has a major influence on choice for class 1, Quality for class 2, and both Fashion and Quality affect the choice for class 3. The small differences in price effect across the three classes turned out to be insignificant.

In addition to the conditional logit model which shows how the attributes affect the likelihood of choosing one alternative over another, differentially for each class, I specified a second logit model to describe the latent class variable as a function of the covariates sex and age. Females turn out to belong more often to class 1 and males to class 3. Younger persons have a higher probability of belonging to class 1 (emphasize Fashion in choices) and older persons are most likely to belong to class 2 (emphasize Quality in choices).

4 LC Models for Nested Data Structures

As explained in the context of the LC regression model, LC analysis is a technique for analyzing two-level data structures. In most cases, this will be repeated measures or item responses that are nested within individuals. Here, I will present a three-level extension of the LC model and discuss the complications in parameter estimation, as well as indicate how these complications can be resolved.

Before proceeding, some additional notation has to introduced. Let y_{ijk} denote the response of individual j within group i on indicator or item k. The number of groups is denoted by N, the number of individuals within group i by n_i, and the number of items by K. The latent class variable at the individual level is denoted as x_j. For reasons that will be clear below, I will use the index j in x when referring to the latent class membership of a certain individual within a group.

The standard method for analyzing such grouped data structures is the multiple-group LC model (see [3]). A multiple-group LC model with group-specific class sizes would be of the form

$$P(\mathbf{y}_i) = \prod_{j=1}^{n_i} \sum_{x=1}^{T} \left\{ \prod_{k=1}^{K} P(y_{ijk}|x) \right\} P(x_j|i).$$

As can be seen, observations within a group are assumed to be independent of each other given the group-specific latent distribution $P(x_j|i)$.

A disadvantage of this "fixed-effects "approach is that the number of unknown parameters increases rapidly as the number of groups increases. An alternative is to assumed that groups belong to latent classes of groups, denoted by w, that differ with respect to the latent distribution of individuals. This yields a LC model of the form

$$P(\mathbf{y}_i) = \sum_{w=1}^{M} \left[\prod_{j=1}^{n_i} \sum_{x_j=1}^{T} \left\{ \prod_{k=1}^{K} P(y_{ijk}|x_j) \right\} P(x_j|w) \right] P(w).$$

This model can be represented as a graphical model containing one latent variable at the group level and one latent variable for each individual within a group. The fact that the model contains so many latent variables makes the use of a standard EM algorithm for maximum likelihood estimation impractical.

The contribution of group i to the completed data log-likelihood that has to be solved in the M step of the EM algorithm has the form

$$\begin{aligned} \log L_i = & \sum_{w=1}^{M} \sum_{x=1}^{T} \sum_{j=1}^{n_i} \sum_{k=1}^{K} P(x_j, w|\mathbf{y}_i) \log P(y_{ijk}|x_j) \\ & + \sum_{w=1}^{M} \sum_{x=1}^{T} \sum_{j=1}^{n_i} P(x_j, w|\mathbf{y}_i) \log P(x_j|w) \\ & + \sum_{w=1}^{M} P(w|\mathbf{y}_i) \log P(w). \end{aligned}$$

This shows that the "only" thing that has to be obtained in the E step of the EM algorithm are the $T \cdot M$ marginal posteriors $P(x_j, w|\mathbf{y}_i)$ for each individual within a group. It turns out that these can be obtained in an efficient manner by making use of the conditional independence assumptions implied by underlying graphical model. More precisely, the new algorithm makes use of the fact that lower-level observations are independent of each other given the higher-level class memberships. The underlying idea of using the structure of the model of interest for the implementation of the EM algorithm is similar to what is done in hidden Markov models. For these models, Baum et al. in [2] developed an efficient EM algorithm which is known as the forward-backward algorithm because it moves forward and backward through the Markov chain. Vermunt in [18] called the version of EM for the new LC model the upward-downward algorithm because it moves upward and downward through the hierarchical structure: First, one marginalizes over class memberships going from the lower to the higher levels. Subsequently, the relevant marginal posterior probabilities are computed going from the higher to the lower levels. The method can easily be generalized to data structures consisting of more than three levels. Moreover, it cannot only be used in LC cluster-like applications, but also in the context of LC regression analysis.

The upward-downward algorithm makes use of the fact that

$$P(x_j, w|\mathbf{y}_i) = P(w|\mathbf{y}_i)P(x_j|\mathbf{y}_i, w) = P(w|\mathbf{y}_i)P(x_j|\mathbf{y}_{ij}, w);$$

that is, that given class membership of the group (w), class membership of the individuals (x_j) is independent of the information of the other group members. The terms $P(w|\mathbf{y}_i)$ and $P(x_j|\mathbf{y}_{ij}, w)$ are obtained with the model parameters:

$$P(x_j|\mathbf{y}_{ij}, w) = \frac{P(x_j, \mathbf{y}_{ij}|w)}{P(\mathbf{y}_{ij}|w)} = \frac{P(x_j|w)\prod_{k=1}^{K} P(y_{ijk}|x_j)}{P(\mathbf{y}_{ij}|w)}$$
$$P(w|\mathbf{y}_i) = \frac{P(w)\prod_{j=1}^{n_i} P(\mathbf{y}_{ij}|w)}{\sum_{w=1}^{M} P(w)\prod_{j=1}^{n_i} P(\mathbf{y}_{ij}|w)},$$

where $P(\mathbf{y}_{ij}|w) = \sum_{x=1}^{T} P(x_j|w)\prod_{k=1}^{K} P(y_{ijk}|x_j)$. In the upward part, we compute $P(x_j, \mathbf{y}_{ij}|w)$ for each individual, collapse these over x_j to obtain $P(\mathbf{y}_{ij}|w)$, and use these to obtain $P(w|\mathbf{y}_i)$ for each group. The downward part involves computing $P(x_j, w|\mathbf{y}_i)$ for each individual using $P(w|\mathbf{y}_i)$ and $P(x_j|\mathbf{y}_{ij}, w)$.

In the upward-downward algorithm computation time increases linearly with the number of individuals within groups instead of exponentially, as would be the case in a standard E step. Computation time can be decreased somewhat more by grouping records with the same values for the observed variables within groups. A practical problem in the implementation of the upward-downward method is that underflows may occur in the computation of $P(w|\mathbf{y}_i)$. More precisely, because it may involve multiplication of a large number $(1 + n_i \cdot K)$ of probabilities, the term $P(w)\prod_{j=1}^{n_i} P(\mathbf{y}_{ij}|w)$ may become equal to zero for each w. Such underflows can, however, easily be prevented by working on a log scale. Letting $a_{iw} = \log[P(w)] + \sum_{j}^{n_i} \log[P(\mathbf{y}_{ij}|w)]$ and $b_i = \max(a_{iw})$, $P(w|\mathbf{y}_i)$ can be obtained as follows:

$$P(w|\mathbf{y}_i) = \frac{\exp\left[a_{iw} - b_i\right]}{\sum_{w}^{M} \exp\left[a_{iw} - b_i)\right]}.$$

4.1 Application: Team Differences in Perceived Task Variety

In a Dutch study on the effect of autonomous teams on individual work conditions, data were collected from 41 teams of two organizations, a nursing home and a domiciliary care organization. These teams contained 886 employees. For the example, I took five dichotomized items of a scale measuring perceived task variety (see [16]). The item wording is as follows (translated from Dutch):

1. Do you always do the same things in your work?
2. Does your work require creativity?
3. Is your work diverse?
4. Does your work make enough usage of your skills and capacities?
5. Is there enough variation in your work?

The original items contained four answer categories. In order simplify the analysis, I collapsed the first two and the last two categories. Because some

Table 5. Test results for the estimated models with the task variety data

Model	Individuals	Groups	Log-likelihood	# parameters	BIC
I	1 class	1 class	-2685	5	5405
II	2 classes	1 class	-2385	11	4844
III	3 classes	1 classes	-2375	16	4859
IV	2 classes	2 classes	-2367	13	4822
V	2 classes	3 classes	-2366	15	4835

respondents had missing values on one or more of the indicators, the estimation procedure was adapted to deal with such partially observed indicators.

The fact that this data set is analyzed by means of LC analysis means that it is assumed that the researcher is interested in building a typology of employees based on their perceived task variety. On other hand, if one would be interested in constructing a continuous scale, a latent trait analysis would be more appropriate. Of course, also in that situation the multilevel structure should be taken into account.

Table 5 reports the log-likelihood value, the number of parameters, and the BIC value for the models that were estimated. I first estimated models without taking the group structure into account. The BIC values for the one to three class model (Models I-III) without a random latent class distribution show that a solution with two classes suffices. Subsequently, I introduced group-specific latent distributions in the two-class model (Models IV and V). From the results obtained with these two models, it can be seen that there is clear evidence for between-team variation in the latent distribution: These models have much lower BIC values than the two-class model without group-specific class sizes. The model with three classes of groups (Model V) has almost the same log-likelihood value as Model IV, which indicates that no more than two latent classes of teams can be identified.

The conditional response probabilities obtained with Model IV indicated that the first class has a much lower probability of giving the high task-variety response than class two on each of the five indicators. The two classes of team members can therefore be named "low task-variety" and "high task-variety". The two classes of teams contained 37 and 63 percent of the teams. The proportion of team members belonging to the high task-variety class are .41 and .78, respectively. This means, for instance, that in the majority of teams (63%) the majority of individuals (78%) belong to the high task-variety group. The substantive conclusion based on Model IV would be that there are two types of employees and two types of teams. The two types of teams differ considerably with respect to the distribution of the team members over the two types of employees.

References

1. Agresti, A.: Categorical Data Analysis. Second Edition, New York: Wiley (2002)

2. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics **41** (1970) 164-171
3. Clogg, C.C., Goodman, L.A.: Latent structure analysis of a set of multidimensional contingency tables. Journal of the American Statistical Association **79** (1984) 762-771.
4. Formann, A.K.: Linear logistic latent class analysis for polytomous data. Journal of the American Statistical Association **87** (1992) 476-486
5. Goldstein, H.: Multilevel statistical models. New York: Halsted Press (1995)
6. Goodman, L.A.: The analysis of systems of qualitative variables when some of the variables are unobservable: Part I - A modified latent structure approach. American Journal of Sociology **79** (1974) 1179-1259
7. Hagenaars, J.A.: Loglinear Models with Latent Variables, Sage University Paper. Newbury Park: Sage Publications (1993)
8. Hagenaars. J.A., McCutcheon., A.L.: Applied Latent Class Analysis, Cambridge: Cambridge University Press (2002)
9. Heinen, T.: Latent Class and Discrete Latent Trait Models: Similarities and Differences. Thousand Oakes: Sage Publications (1996)
10. Kamakura, W.A., Wedel, M., Agrawal, J.: Concomitant variable latent class models for the external analysis of choice data. International Journal of Marketing Research **11** (1994) 541-464
11. Lazarsfeld, P.F., Henry, N.W.: Latent Structure Analysis. Boston: Houghton Mill (1968)
12. Magidson, J., Vermunt, J.K.: Latent class factor and cluster models, bi-plots and related graphical displays, Sociological Methodology **31** (2001) 223-264
13. McCutcheon, A.L.: Latent Class Analysis, Sage University Paper. Newbury Park: Sage Publications (1987)
14. McGrath, K., Waterton, J.: British social attitudes, 1983-1986 panel survey. London: Social and Community Planning Research, Technical Report (1986)
15. McLachlan, G.J., Peel, D.: Finite Mixture models. New York: John Wiley & Sons, Inc. (2000)
16. Van Mierlo, H., Vermunt, J.K., Rutte, C.: Using individual level survey data to measure group constructs: A comparison of items with reference to the individual and to the group in a job design context (submitted for publication)
17. Vermunt, J.K.: Log-linear Models for Event Histories. Thousand Oakes: Series QASS, vol 8. Sage Publications (1997)
18. Vermunt, J.K.: Multilevel latent class models, Sociological Methodology, 33, (to appear)
19. Vermunt, J.K.: The use restricted latent class models for defining and testing nonparametric and parametric IRT models. Applied Psychological Measurement **25** (2001) 283-294
20. Vermunt, J.K., Magidson, J.: Latent GOLD 2.0 User's Guide. Belmont, MA: Statistical Innovations Inc. (2000)
21. Vermunt, J.K., Magidson, J.: Addendum to the Latent GOLD User's Guide: Upgrade Manual for Version 3.0. Belmont, MA: Statistical Innovations Inc. (2003)
22. Vermunt J.K., Magidson, J.: Factor Analysis with Categorical Indicators: A Comparison Between Traditional and Latent Class Approaches. A. Van der Ark, M. Croon, and K.Sijstma. Advancements in Categorical Data Analysis. Erlbaum, (to appear)
23. Vermunt, J.K., Van Dijk. L.: A nonparametric random-coefficients approach: the latent class regression model. Multilevel Modelling Newsletter **13** (2001) 6-13

24. Wedel, M. and DeSarbo, W.: Mixture regression models. J. Hagenaars and A. McCutcheon (eds.), Applied Latent Class Analysis, 366- 382. Cambridge University Press (2002)
25. Zhang, N.L.: Hierarchical latent class models for cluster analysis. Journal of Machine Learning Research (to appear)