Latent Class Analysis

Jeroen K. Vermunt & Jay Magidson

The basic idea underlying latent class (LC) analysis is a very simple one: some of the parameters of a postulated statistical model differ across unobserved subgroups. These subgroups form the categories of a categorical latent variable (see entry LATENT VARIABLE). This basic idea has several seemingly unrelated applications, the most important of which are clustering, scaling, density estimation, and random-effects modeling. Outside social sciences, LC models are often referred to as finite mixture models.

LC analysis was introduced in 1950 by Lazarsfeld, who used the technique as a tool for building typologies (or clustering) based on dichotomous observed variables. More than 20 years later, Goodman (1974) made the model applicable in practice by developing an algorithm for obtaining maximum likelihood estimates of the model parameters. He also proposed extensions for polytomous manifest variables and multiple latent variables, and did important work on the issue of model identification. During the same period, Haberman (1979) showed the connection between LC models and log-linear models for frequency tables with missing (unknown) cell counts. Many important extensions of the classical LC model have been proposed since then, such as models containing (continuous) covariates, local dependencies, ordinal variables, several latent variables, and repeated measures. A general framework for categorical data analysis with discrete latent variables was proposed by Hagenaars (1990) and extended by Vermunt (1997).

While in the social sciences LC and finite mixture models are conceived primarily as tools for categorical data analysis, they can be useful in several other areas as well. One of these is density estimation, in which one makes use of the fact that a complicated density can be approximated as a finite mixture of simpler densities. LC analysis can also be used 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 (see LATENT PROFILE MODEL). Another application area is dealing with unobserved heterogeneity, for example, in regression analysis with dependent observations (see NON-PARAMETRIC RANDOM-EFFECTS MODEL).

The classical LC model for categorical indicators

Let X represent the latent variable and Y_ℓ one of the L observed or manifest variables, where $1 \leq \ell \leq L$. Moreover, let C be the number of latent classes and D_ℓ the number of levels of Y_ℓ. A particular LC is enumerated by the index x, $x = 1, 2, ..., C$, and a particular value of Y_ℓ by y_ℓ, $y_\ell = 1, 2, ..., D_\ell$. The vector notation $\mathbf{Y}$ and $\mathbf{y}$ is used to refer to a complete response pattern.

In order to make things more concrete, consider the following small data set obtained from the 1987 General Social Survey:

Y_1	Y_2	Y_3	Frequency	$P(X = 1\|\mathbf{Y} = \mathbf{y})$	$P(X = 2\|\mathbf{Y} = \mathbf{y})$
1	1	1	696	.998	.002
1	1	2	68	.929	.071
1	2	1	275	.876	.124
1	2	2	130	.168	.832
2	1	1	34	.848	.152
2	1	2	19	.138	.862
2	2	1	125	.080	.920
2	2	2	366	.002	.998

The three dichotomous indicators Y_1, Y_2, and Y_3 are the responses to the statements "allow anti-religionists to speak"(1 = allowed, 2 = not allowed), "allow anti-religionists to teach" (1 = allowed, 2 = not allowed), "remove anti-religious books from the library" (1 = do not remove, 2 = remove). By means of LC analysis it is possible to identify subgroups with different degrees of tolerance towards anti-religionists.

The basic idea underlying any type of LC model is that the probability of obtaining response pattern $\mathbf{y}$, $P(\mathbf{Y} = \mathbf{y})$, is a weighted average of the C class-specific probabilities $P(\mathbf{Y} = \mathbf{y}|X = x)$; that is,

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} P(X = x)P(\mathbf{Y} = \mathbf{y}|X = x). \tag{1}$$

Here, $P(X = x)$ denotes the proportion of persons belonging to LC x.

In the classical LC model, this basic idea is combined with the assumption of LOCAL INDEPENDENCE. The L manifest variables are assumed to be

mutually independent within each LC, which can be formulated as follows:

$$P(\mathbf{Y} = \mathbf{y}|X = x) = \prod_{\ell=1}^{L} P(Y_\ell = y_\ell|X = x). \tag{2}$$

After estimating the conditional response probabilities $P(Y_\ell = y_\ell|X = x)$, comparing these probabilities between classes shows how the classes differ from each other, which can be used to name the classes. Combining the two basic equations (1) and (2) yields the following model for $P(\mathbf{Y} = \mathbf{y})$:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} P(X = x) \prod_{\ell=1}^{L} P(Y_\ell = y_\ell|X = x).$$

A two-class model estimated with the small example data set yielded the following results:

	$X = 1$ (Tolerant)	$X = 2$ (Intolerant)
$P(X = x)$	.62	.38
$P(Y_1 = 1\|X = x)$	.96	.23
$P(Y_2 = 1\|X = x)$	.74	.04
$P(Y_3 = 1\|X = x)$	.92	.24

The two classes contain 62 and 38 percent of the individuals, respectively. The first class can be named "Tolerant" because people belonging to that class have much higher probabilities of selecting the tolerant responses on the indicators than people belonging to the second "Intolerant" class.

Similarly to cluster analysis, one of the purposes of LC analysis might be to assign individuals to latent classes. The probability of belonging to LC x – often referred to as posterior membership probability – can be obtained by the Bayes rule,

$$P(X = x|\mathbf{Y} = \mathbf{y}) = \frac{P(X = x)P(\mathbf{Y} = \mathbf{y}|X = x)}{P(\mathbf{Y} = \mathbf{y})}. \tag{3}$$

The most common classification rule is modal assignment, which amounts to assigning each individual to the LC with the highest $P(X = x|\mathbf{Y} = \mathbf{y})$. The class-membership probabilities reported in the first table show that people with at least two tolerant responses are classified into the "Tolerant" class.

Log-linear formulation of the LC model

Haberman (1979) showed that the LC model can also be specified as a LOG-LINEAR MODEL for a table with missing cell entries or, more precisely, as a model for the expanded table including the latent variable X as an additional dimension. The relevant log-linear model for $P(X = x, \mathbf{Y} = \mathbf{y})$ has the following form:

$$\ln P(X = x, \mathbf{Y} = \mathbf{y}) = \beta + \beta_x^X + \sum_{\ell=1}^{L} \beta_{y_\ell}^{Y_\ell} + \sum_{\ell=1}^{L} \beta_{x,y_\ell}^{X,Y_\ell}.$$

It contains a main effect, the one-variable terms for the latent variable and the indicators, and the two-variable terms involving X and each of the indicators. Note that the terms involving two or more manifest variables are omitted because of the local independence assumption.

The connection between the log-linear parameters and the conditional response probabilities is as follows:

$$P(Y_\ell = y_\ell | X = x) = \frac{\exp(\beta_{y_\ell}^{Y_\ell} + \beta_{x,y_\ell}^{X,Y_\ell})}{\sum_{r=1}^{D_\ell} \exp(\beta_r^{Y_\ell} + \beta_{x,r}^{X,Y_\ell})}.$$

This shows that the log-linear formulation amounts to specifying a logit model for each of the conditional response probabilities.

The type of LC formulation that is used becomes important if one wishes to impose restrictions. Although constraints on probabilities can sometimes be transformed into constraints on log-linear parameters and vice versa, there are many situations in which this is not possible.

Maximum likelihood estimation

Let I denote the total number of cells entries (or possible answer patterns) in the L-way frequency table, so that $I = \prod_{\ell=1}^{L} D_\ell$, and let i denote a particular cell entry, n_i the observed frequency in cell i, and $P(\mathbf{Y} = \mathbf{y}_i)$ the probability of having the response pattern of cell i.

The parameters of LC models are typically estimated by means of maximum likelihood (ML). The kernel of the log-likelihood function that is maximized equals

$$\ln \mathcal{L} = \sum_{i=1}^{I} n_i \ln P(\mathbf{Y} = \mathbf{y}_i)$$

Notice that only non-zero 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.

One of the problems in the estimation of LC models is that model parameters may be non-identified, even if the number of degrees of freedom is larger or equal to zero. Non-identification 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 information matrix which should be positive definite. 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 at possible pitfalls. For an unrestricted LC analysis, one needs at least three indicators, but if these are dichotomous, no more than two latent classes can be identified. One has to watch out with four dichotomous 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 indicators, however, even a five-class model is identified. Usually, it is possible to achieve identification by constraining certain model parameters: for example, the restrictions $P(Y_\ell = 1|X = 1) = P(Y_\ell = 2|X = 2)$ can be used to identify a two-class model with two dichotomous indicators.

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. Some software packages have automated the use of several sets of random starting values in order 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 zero (or one) or log-linear 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 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 an estimated LC model is usually tested by the Pearson or the likelihood-ratio chi-squared statistic (see CATEGORICAL DATA ANALYSIS). The latter is defined as

$$L^2 = 2\sum_{i=1}^{I} n_i \ln \frac{n_i}{N \cdot P(\mathbf{Y} = \mathbf{y}_i)},$$

where N denotes the total sample size. As in log-linear analysis, the number of degrees of freedom (df) equals the number of cells in the frequency table minus one, $\prod_{\ell=1}^{L} D_\ell - 1$, minus the number of independent parameters. In an unrestricted LC model,

$$df = \prod_{\ell=1}^{L} D_\ell - C \cdot \left[1 + \sum_{\ell=1} (D_\ell - 1)\right].$$

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 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; for example, in the two-way marginal tables.

It is not valid to compare models with C and $C+1$ classes by subtracting their L^2 and df values because this conditional test does not have an asymptotic chi-squared distribution. This means that alternative methods are required for comparing models with different numbers of classes. One popular method is the use of information criteria such as BIC and AIC.

Another more descriptive method is a measure for the proportion of total association accounted for by a C-class model, $[L^2(1) - L^2(C)]/L^2(1)$, where the L^2 value of the one-class (independence) model, $L^2(1)$, is used as a measure of total association in the L-way frequency table.

Usually we are not only interested in goodness-of-fit, but also in the performance of the modal classification rule [see equation (3)]. The estimated proportion of classification errors under modal classification equals

$$E = \sum_{i=1}^{I} \frac{n_i}{N} \left\{1 - \max\left[P(X = x|\mathbf{Y} = \mathbf{y}_i)\right]\right\}.$$

This number can be compared to the proportion of classification errors based on the unconditional probabilities $P(X = x)$, yielding a reduction of errors measure λ:

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

The closer this nominal R^2-type measure is to one, the better the classification performance of a model.

Extensions of the LC model for categorical indicators

Several extensions have been proposed of the basic LC model. One of the most important extensions is the inclusion of covariates or grouping variables which describe (predict) the latent variable X. This is achieved by specifying a multinomial logit model for the probability of belonging to LC x; that is,

$$P(X = x|\mathbf{Z} = \mathbf{z}) = \frac{\exp(\gamma_x^X + \sum_{k=1}^{K} \gamma_x^{X,Z_k} \cdot z_k)}{\sum_{r=1}^{C} \exp(\gamma_r^X + \sum_{k=1}^{K} \gamma_r^{X,Z_k} \cdot z_k)},$$

where z_k denotes a value of covariate k.

Another important extension is related to the use of information on the ordering of categories. Within the log-linear LC framework, ordinal constraints can be imposed via ASSOCIATION MODEL structures for the two-variable terms $\beta_{x,y_\ell}^{X,Y_\ell}$. For example, if Y_ℓ is an ordinal indicator, we can restrict $\beta_{x,y_\ell}^{X,Y_\ell} = \beta_x^{X,Y_\ell} \cdot y_\ell$. Similar constraints can be used for the latent variable (Heinen, 1996).

In the case that a C-class model does not fit the data, the local independence assumption fails to hold for one or more pairs of indicators. The common model fitting strategy in LC analysis is to increase the number of

latent classes till the local independence assumption holds. Two extensions have been developed that make it possible to follow other strategies. Rather than increasing the number of latent classes, one alternative approach is to relax the local independence assumption by including direct effects between certain indicators – a straightforward extension to the log-linear LC model. Another alternative strategy involves increasing the number of latent variables instead of the number of latent classes. This so-called LC factor analysis approach (Magidson and Vermunt, 2001) is especially useful if the indicators measure several dimensions.

Other important extensions involve the analysis of longitudinal data (see LATENT MARKOV MODEL) and partially observed indicators. The most general model that contains all models discussed thus far as special cases is the structural equation model for categorical data proposed by Hagenaars (1990) and Vermunt (1997).

Other types of LC models

Thus far, we have focused on LC models for categorical indicators. However, the basic idea of LC analysis, that parameters of a statistical model differ across unobserved subgroups, can also be applied with variables of other scales types. In particular, there are three important types of applications of LC or finite mixture models that fall outside the categorical data analysis framework: clustering with continuous variables, density estimation, and random-effects modeling.

Over the past ten years, there has been a renewed interest in LC analysis as a tool for cluster analysis with continuous indicators. The LC model can be seen as a probabilistic or model-based variant of traditional non-hierarchical cluster analysis procedures such as the K-means method. It has been shown that such a LC-based clustering procedure outperforms the more ad hoc traditional methods. The method is known under names such as LATENT PROFILE MODEL, mixture-model clustering, model-based clustering, latent discriminant analysis, and LC clustering. The basic formula of this model is similar to one given in equation (1); that is,

$$f(\mathbf{Y}=\mathbf{y}) = \sum_{x=1}^{C} P(X=x) f(\mathbf{Y}=\mathbf{y}|X=x).$$

As shown by this slightly more general formulation, the probabilities $P(...)$ are replaced by densities $f(...)$. With continuous variables, the class-specific

densities $f(\mathbf{Y} = \mathbf{y}|X = x)$ will usually be assumed to be (restricted) multivariate normal, where each LC has its own mean vector and covariance matrix. Note that this is a special case of the more general principle of density estimation by finite mixtures of simple densities.

Another important application of LC analysis is as a NONPARAMETRIC RANDOM-EFFECTS MODEL. The idea underlying this application is that the parameters of the regression model of interest may differ across unobserved subgroups. For this kind of analysis, often referred to as LC regression analysis, the LC variable serves the role of a MODERATING VARIABLE. The method is very similar to regression models for repeated measures or two-level data sets, with the difference that no assumptions are made about the distribution of the random coefficients.

Software

The first LC program, MLLSA, made available by Clifford Clogg in 1977, was limited to a relative small number of nominal variables. Today's program 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 some kinds of LC models, but not for nominal indicators. In contrast to these command language programs, Latent GOLD is a program with an SPSS-like user interface that is especially developed for LC analysis. It implements the most important types of LC models, deals with variables of different scale types, and extends the basic model to include covariates, local dependencies, several latent variables, and partially observed indicators.

References

Goodman, L.A. (1974). 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, 1179-1259.

Haberman, S.J. (1979). *Analysis of Qualitative Data, Vol 2, New Developments*. New York: Academic Press.

Hagenaars, J.A. (1990). *Categorical Longitudinal Data - Loglinear Analysis of Panel, Trend and Cohort Data*. Newbury Park: Sage.

Hagenaars, J.A. and McCutcheon, A.L. (2002), *Applied Latent Class Analysis*. Cambridge University Press.

Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Thousand Oakes: Sage Publications.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis & The interpretation and mathematical foundation of latent structure analysis. S.A. Stouffer et al. (eds.), *Measurement and Prediction*, 362-472. Princeton, NJ: Princeton University Press.

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

Vermunt, J.K. (1997). *Log-linear Models for Event Histories*. Thousand Oakes: Sage Publications.