

Latent Class Models

by

Jay Magidson, Ph.D.
Statistical Innovations Inc.

Jeroen K. Vermunt, Ph.D.
Tilburg University, the Netherlands

Over the past several years more significant books have been published on latent class and other types of finite mixture models than any other class of statistical models. The recent increase in interest in latent class models is due to the development of extended computer algorithms, which allow today's computers to perform latent class analysis on data containing more than just a few variables. In addition, researchers are realizing that the use of latent class models can yield powerful improvements over traditional approaches to cluster, factor, regression/segmentation, as well as to multivariable biplots and related graphical displays.

What are Latent Class Models?

Traditional models used in regression, discriminant and log-linear analysis contain parameters that describe only relationships between the observed variables. Latent class (LC) models (also known as finite mixture models) differ from these by including one or more discrete unobserved variables. In the context of marketing research, one will typically interpret the categories of these latent variables, the latent classes, as clusters or segments (Dillon and Kumar 1994; Wedel and Kamakura 1998). In fact, LC analysis provides a powerful new tool to identify important market segments in target marketing.

LC models do not rely on the traditional modeling assumptions which are often violated in practice (linear relationship, normal distribution, homogeneity). Hence, they are less subject to biases associated with data not conforming to model assumptions. In addition, LC models have recently been extended (Vermunt and Magidson, 2000a, 2000b) to include variables of mixed scale types (nominal, ordinal, continuous and/or count variables) in the same analysis. Also, for improved cluster or segment description the relationship between the latent classes and external variables (covariates) can be assessed simultaneously with the identification of the clusters. This eliminates the need for the usual second stage of analysis where a discriminant analysis is performed to relate the cluster results to demographic and other variables.

Kinds of Latent Class Models

Three common statistical application areas of LC analysis are those that involve

- 1) clustering of cases,
- 2) variable reduction and scale construction, and
- 3) prediction.

This paper introduces the three major kinds of LC models:

- LC Cluster Models,
- LC Factor Models,
- LC Regression Models.

Our illustrative examples make use of the new computer program (Vermunt and Magidson, 2000b) called Latent GOLD[®].

LC Cluster Models

The **LC Cluster** model:

- identifies *clusters* which group together *persons* (cases) who share similar interests/values/characteristics/behavior,
- includes a K-category latent variable, each category representing a cluster.

Advantages over traditional types of cluster analysis include:

- probability-based classification: Cases are classified into clusters based upon membership probabilities estimated directly from the model,
- variables may be continuous, categorical (nominal or ordinal), or counts or any combination of these,
- demographics and other covariates can be used for cluster description.

Typical marketing applications include:

- exploratory data analysis,
- development of a behavioral based and other segmentations of customers and prospects.

Traditional clustering approaches utilize unsupervised classification algorithms that group cases together that are "near" each other according to some ad hoc definition of "distance". In the last decade interest has shifted towards model-based approaches which use estimated membership probabilities to classify cases into the appropriate cluster. The most popular model-based approach is known as mixture-model clustering, where each latent class represents a hidden cluster (McLachlan and Basford, 1988). Within the marketing research field, this method is sometimes referred to as "latent discriminant analysis" (Dillon and Mulani, 1999). Today's high-speed computers make these computationally intensive methods practical.

For the general finite mixture model, not only continuous variables, but also variables that are ordinal, nominal or counts, or any combination of these can be included. Also, covariates can be included for improved cluster description.

As an example, we used the LC cluster model to develop a segmentation of current bank customers based upon the types of accounts they have. Separate models were developed specifying different numbers of clusters and the model selected was the one that had the lowest BIC statistic.

This criteria resulted in 4 segments which were named:

- 1) Value Seekers (15% of customers),
- 2) Conservative Savers (35% of customers),
- 3) Mainstreamers (40% of customers),
- 4) Investors (10% of customers).

For each customer, the model gave estimated membership probabilities for each segment based on their account mix. The resulting segments were verified to be very homogeneous and to differ substantially from each other not only with respect to their mix of accounts, but also with respect to demographics, and profitability. In addition, examination of survey data among the sample of customers for which customer satisfaction data were obtained found some important attitudinal and satisfaction differences between the segments as well. Value seekers were youngest and a high percentage were new customers. Basic savers were oldest.

Investors were the most profitable customer segment by far. Although only 10% of all customers, they accounted for over 30% of the bank's deposits. Survey data pinpointed the areas of the bank with which this segment was least satisfied and a LC regression model (see below) on follow-up data related their dissatisfaction to attrition. The primary uses of the survey data was to identify reasons for low satisfaction and to develop strategies of improving satisfaction in the manner that increased retention.

This methodology of segmenting based on behavioral information available on all customers offers many advantages over the common practice of developing segments from survey data and then attempting to allocate all customers to the different clusters.

Advantages of developing a segmentation based on *behavioral* data include:

- past behavior is known to be the best predictor of future behavior,
- all customers can be assigned to a segment directly, not just the sample for which survey data is available,
- improved reliability over segmentations based on attitudes, demographics, purchase intent and other survey variables (when segment membership is based on survey data, a large amount of classification error is almost always present for non-surveyed customers) .

LC Factor Models

The **LC Factor** model:

- identifies *factors* which group together *variables* sharing a common source of variation,
- can include several ordinal latent variables, each of which contains 2 or more levels,
- is similar to maximum likelihood factor analysis in that its use may be exploratory or confirmatory and factors may be assumed to be correlated or uncorrelated (orthogonal).

Advantages over traditional factor analysis are:

- factors need not be rotated to be interpretable,
- factor scores are obtained directly from the model without imposing additional assumptions,
- variables may be continuous, categorical (nominal or ordinal), or counts or any combination of these,
- extended factor models can be estimated that include covariates and correlated residuals.

Typical marketing applications include:

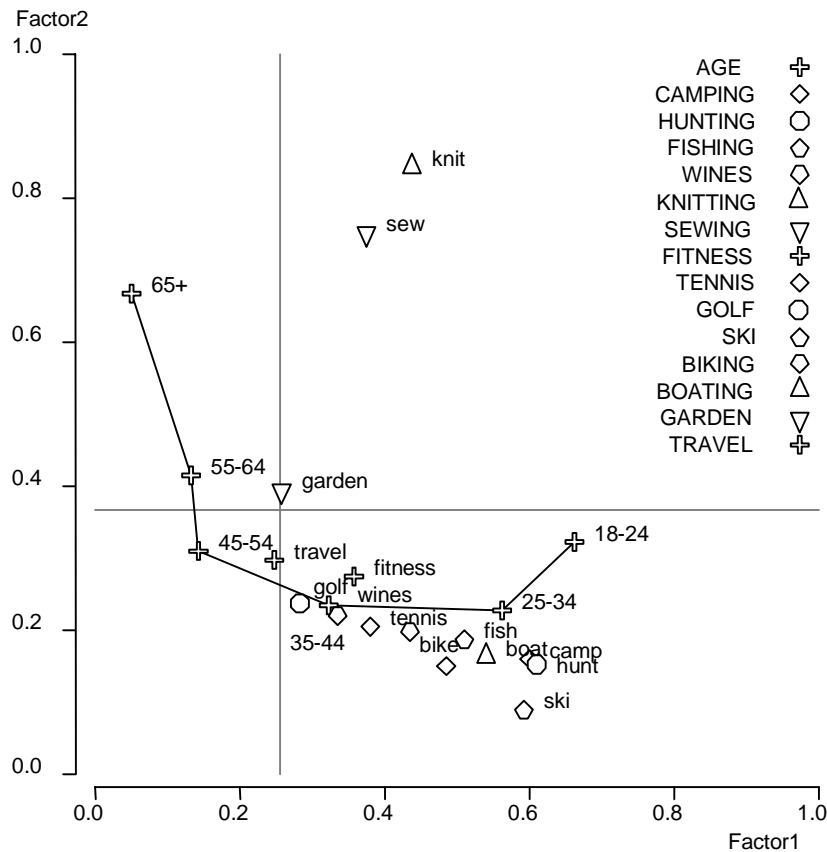
- development of composite variables from attitudinal survey items,
- development of perceptual maps and other kinds of biplots which relate product and brand usage to behavioral and attitudinal measures and to demographics,
- estimation of factor scores,
- direct conversion from factors to segments.

The conversion of ordinal factors to segments is straightforward. For example, consider a model containing 2 dichotomous factors. In this case, the LC factor model provides membership classification probabilities directly for 4 clusters (segments) based on the classification of cases as high vs. low on each factor: segment 1 = (low, low); segment 2 = (low, high); segment 3 = (high, low) and segment 4 = (high, high). Magidson and Vermunt (2000) found that LC factor models specifying uncorrelated factors often fit data better than comparable cluster models (i.e., cluster models containing the same number of parameters).

Figure 1 provides a bi-plot in 2-factor space of lifestyle interests where the horizontal axis represents the probability of being high on factor 1 and the vertical axis the probability of being high on factor 2. The variable AGE was included directly in the LC Factor model as a covariate and therefore shows up in the bi-plot to assist in understanding the meaning of the factors. For example, we see that persons aged 65+ are most likely to be in the (low, high) segment, as are persons expressing an interest in sewing. As a group, their (mean) factor scores are (Factor 1, Factor 2) = (.06, .67).

Since these factor scores have a distinct probabilistic interpretation, this bi-plot represents an improvement over traditional biplots and perceptual maps (see Magidson and Vermunt 2000). Individual cases can also be plotted based on their factor scores.

Figure 1: Bi-plot for life-style data



The factor model can also be used to deal with measurement and classification errors in categorical variables. It is actually equivalent to a latent trait (IRT) model without the requirement that the traits be normally distributed.

LC Regression Models

The **LC Regression** model, also known as the **LC Segmentation** model:

- is used to predict a dependent variable as a function of predictors,
- includes an R-category latent variable, each category representing a homogeneous population (class, segment),
- different regressions are estimated for each population (for each latent segment),
- classifies cases into segments and develops regression models simultaneously.

Advantages over traditional regression models include:

- relaxing the traditional assumption that the same model holds for all cases ($R=1$) allows the development of separate regressions to be used to target each segment,
- diagnostic statistics are available to determine the value for R,

- for $R > 1$, covariates can be included in the model to improve classification of each case into the most likely segment.

Typical marketing applications include:

- customer satisfaction studies: identify particular determinants of customer satisfaction that are appropriate for each customer segment,
- conjoint studies: identify the mix of product attributes that appeal to different market segments,
- more generally: identify segments that differ from each other with respect to some dependent variable criterion.

Like traditional regression modeling, LC regression requires a computer program. As LC regression modeling is relatively new, very few programs currently exist. Our comparisons between LC regression and traditional linear regression are based on the particular forms of LC regression that are implemented in the Latent GOLD[®] program. For other software see Wedel and DeSarbo (1994) and Wedel and Kamakura (1998).

Typical regression programs utilize ordinary least squares estimation in conjunction with a linear model. In particular, such programs are based on two restrictive assumptions about data that are often violated in practice:

- 1) the dependent variable is continuous with prediction error normally distributed,
- 2) the population is homogeneous - one model holds for all cases.

LC regression as implemented in the Latent GOLD[®] program relaxes these assumptions:

- 1) it accommodates dependent variables that are continuous, categorical (binary, polytomous nominal or ordinal), binomial counts, or Poisson counts,
- 2) the population needs not be homogeneous (i.e., there may be multiple populations as determined by the BIC statistic).

One potential drawback for LC models is that there is no guarantee that the solution will be the maximum likelihood solution. LC computer programs typically employ the EM or Newton Raphson algorithm which may converge to a local as opposed to a global maximum. Some programs provide randomized starting values to allow users to increase the likelihood of converging to a global solution by starting the algorithm at different randomly generated starting places. An additional approach is to use Bayesian prior information in conjunction with randomized starting values which eliminates the possibility of obtaining boundary (extreme) solutions and reduces the chance of obtaining local solutions. Generally speaking, we have achieved good results using 10 randomized starting values and small Bayes constants (the default option in the Latent GOLD program).

In addition to using predictors to estimate separate regression model for each class, covariates can be specified to refine class descriptions and improve classification of cases into the appropriate latent classes. In this case, LC regression analysis consists of 3 simultaneous steps:

- 1) identify latent classes or hidden segments

- 2) use demographic and other covariates to predict class membership, and
- 3) classify cases into the appropriate classes/segments

Dependent variables may also include repeated/correlated observations of the kind often collected in conjoint marketing studies where persons are asked to rate different product profiles. Below is an example of a full factorial conjoint study designed to assist in the determination of the mix of product attributes for a new product.

Conjoint Case Study

In this example, 400 persons were asked to rate each of 8 different attribute combinations regarding their likelihood to purchase. Hence, there are 8 records per case; one record for each cell in this 2x2x2 conjoint design based on the following attributes:

- *FASHION* (1 = Traditional; 2 = Modern),
- *QUALITY* (1 = Low; 2 = High),
- *PRICE* (1 = Lower; 2 = Higher) .

The dependent variable (*RATING*) is the rating of purchase intent on a five-point scale. The three attributes listed above are used as predictor variables in the model and the following demographic variables are used as covariates:

- *SEX* (1 = Male; 2 = Female),
- *AGE* (1 = 16-24; 2 = 25-39; 3 = 40+).

The goal of a traditional conjoint study of this kind is to determine the relative effects of each attribute in influencing one's purchase decision; a goal attained by estimating regression (or logit) coefficients for these attributes. When the LC regression model is used with the same data, a more general goal is attained. First, it is determined whether the population is homogeneous or whether there exists two or more distinct populations (latent segments) which differ with respect to the relative importance placed on each of the three attributes. If multiple segments are found, separate regression models are estimated simultaneously for each. For example, for one segment, price may be found to influence the purchase decision, while a second segment may be price insensitive, but influenced by quality and modern appearance.

We will treat *RATING* as an *ordinal* dependent variable and estimate several different models to determine the number of segments (latent classes). We will then show how this methodology can be used to describe the demographic differences between these segments and to classify each respondent into the segment which is most appropriate.

We estimated one- to four-class models with and without covariates. Table 1 reports the obtained test results. The BIC values indicate that the three-class model is the best model (BIC is lowest for this model) and that the inclusion of covariates significantly improves the model.

Table 1: Test results for regression models for conjoint data

Model	Log-likelihood	BIC-value	Number of parameters
Without covariates			
One segment	-4402	8846	7
Two segments	-4141	8319	15
Three segments	-4087	8312	23
Four segments	-4080	8346	31
With covariates			
Two segments	-4088	8284	18
Three segments	-4036	8246	29
Four segments	-4026	8293	40

The parameter estimates of the three-class model with covariates are reported in Tables 2 and 3 and 4. As can be seen from the first row of Table 2, segment 1 contains about 50% of the subjects, segment 2 contains about 25% and segment 3 contains the remaining 25%. Examination of class-specific probabilities shows that overall, segment 1 is least likely to buy (only 5% are *Very Likely* to buy) and segment 3 is most likely (21% are *Very Likely* to buy).

◆ **Table 2: Profile output**

	Class 1	Class 2	Class 3
Segment Size	0.49	0.26	0.25
Rating			
Very Unlikely	0.21	0.10	0.05
Not Very Likely	0.43	0.20	0.12
Neutral	0.20	0.37	0.20
Somewhat Likely	0.10	0.21	0.43
Very Likely	0.05	0.11	0.21

◆ **Table 3: Beta's or parameters of model for dependent variable**

	Class 1	Class 2	Class 3	Wald	p-value	Wald(=)	p-value
Fashion	1.97	1.14	0.04	440.19	4.4e-95	191.21	3.0e-42
Quality	0.04	0.85	2.06	176.00	6.5e-38	132.33	1.8e-29
Price	-1.04	-0.99	-0.94	496.38	2.9e-107	0.76	0.68

The beta parameter for each predictor is a measure of the influence of that predictor on *RATING*. The beta effect estimates under the column labeled Class 1 suggest that segment 1 is influenced in a positive way by products for which *FASHION* = *Modern* (beta = 1.97) and in negative way by *PRICE* = *Higher* (beta = -1.04), but not by *QUALITY* (beta is approximately 0). We also see that segment 2 is influenced by all 3 attributes, having a preference for those product choices that are *modern* (beta = 1.14), *high quality* (beta = .85) and *lower priced* (beta = -0.99). Members of segment 3 prefer

high quality (beta = 2.06) and the *lower* (beta = -.94) product choices, but are not influenced by *FASHION*.

Note that *PRICE* has more or less the same influence on all three segments. The Wald (=) statistic indicates that the differences in these beta effects across classes are not significant (the *p*-value = .68 which is much higher than .05, the standard level for assessing statistical significance). This means that all 3 segments exhibit price sensitivity to the same degree. This is confirmed when we estimate a model in which this effect is specified to be class-independent. The *p*-value for the Wald statistic for *PRICE* is 2.9×10^{-107} indicating that the amount of price sensitivity is highly significant.

With respect to the effect of the other two attributes we find large between-segment differences. The predictor *FASHION* has a strong influence on segment 1, a less strong effect on segment 2, and virtually no effect on segment 3. *QUALITY* has a strong effect on segment 3, a less strong effect on segment 2, and virtually no effect on segment 1. The fact that the influence of *FASHION* and *QUALITY* differs significantly between the 3 segments is confirmed by the significant *p*-values associated with the Wald(=) statistics for these attributes. For example, for *FASHION*, the *p*-value = 3.0×10^{-42} .

The beta parameters of the regression model can be used to name the latent segments. Segment 1 could be named the “Fashion-Oriented” segment, segment 3 the “Quality-Oriented” segment, and segment 2 is the segment that takes into account all 3 attributes in their purchase decision.

◆ **Table 4: Gamma's: parameters of model for latent distribution**

	Class 1	Class 2	Class 3	Wald	p-value
Sex					
Male	-0.56	0.71	-0.15	24.47	4.9e-6
Female	0.56	-0.71	0.15		
Age					
16-25	0.84	-0.59	-0.24	53.09	8.1e-11
26-40	-0.32	0.59	-0.27		
40+	-0.52	0.01	0.51		

The parameters of the (multinomial logit) model for the latent distribution appear in Table 4. These show that females have a higher probability of belonging to the “Fashion-oriented” segment (segment 1), while males more often belong to segment 2. The Age effects show that the youngest age group is over-represented in the “Fashion-oriented” segment, while the oldest age group is over-represented in the “Quality oriented” Segment.

Conclusions

We introduced three kinds of LC models and described applications of each that are of interest in marketing research, survey analysis and related fields. It was shown that LC analysis can be used as a replacement for traditional cluster analysis techniques, as a factor analytic tool for reducing dimensionality, and as a tool for estimating separate regression models for each segment. In particular, these models offer powerful new approaches for identifying market segments.

BIOS

Jay Magidson is founder and president of Statistical Innovations, a Boston based consulting, training and software development firm specializing in segmentation modeling. His clients have included A.C. Nielsen, Household Finance, and National Geographic Society. He is widely published on the theory and applications of multivariate statistical methods, and was awarded a patent for a new innovative graphical approach for analysis of categorical data. He taught statistics at Tufts and Boston University, and is chair of the Statistical Modeling Week workshop series. Dr. Magidson designed the SPSS CHAID™ and GOLDMineR® programs, and is the co-developer (with Jeroen Vermunt) of Latent GOLD®.

Jeroen Vermunt is Assistant Professor in the Methodology Department of the Faculty of Social and Behavioral Sciences, and Research Associate at the Work and Organization Research Center at Tilburg University in the Netherlands. He has taught a variety of courses and seminars on log-linear analysis, latent class analysis, item response models, models for non-response, and event history analysis all over the world, as well as published extensively on these subjects. Professor Vermunt is developer of the LEM program and co-developer (with Jay Magidson) of Latent GOLD®.

References

- Dillon, W.R., and Kumar, A. (1994). Latent structure and other mixture models in marketing: An integrative survey and overview. R.P. Bagozzi (ed.), *Advanced methods of Marketing Research*, 352-388, Cambridge: Blackwell Publishers.
- Dillon, W.R.. and Mulani, N. (1989) LADI: A latent discriminant model for analyzing marketing research data. *Journal of Marketing Research*, 26, 15-29.
- Magidson J., and Vermunt, J.K. (2000), *Latent Class Factor and Cluster Models, Bi-plots and Related Graphical Displays*. Submitted for publication.
- McLachlan, G.J., and Basford, K.E. (1988). *Mixture models: inference and application to clustering*. New York: Marcel Dekker.
- Vermunt, J.K. & Magidson, J. (2000a). "Latent Class Cluster Analysis", chapter 3 in J.A. Hagenaars and A.L. McCutcheon (eds.), *Advances in Latent Class Analysis*. Cambridge University Press.
- Vermunt, J.K. & Magidson, J. (2000b). *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.
- Wedel, M., and DeSarbo, W.S (1994). A review of recent developments in latent class regression models. R.P. Bagozzi (ed.), *Advanced methods of Marketing Research*, 352-388, Cambridge: Blackwell Publishers.
- Wedel, M., and Kamakura, W.A. (1998). *Market segmentation: Concepts and methodological foundations*. Boston: Kluwer Academic Publishers.

