New Developments in Latent Class Choice Models

by
Jay Magidson, Statistical Innovations Inc.
Tom Eagle, Eagle Analytics
Jeroen K. Vermunt, Tilburg University

Discrete choice models have proven to be good methods for predicting market shares for new products based on consumers' expressed preferences between choice alternatives. However, the standard aggregate model fails to take into account the fact that preferences (utilities) differ from one respondent to another (or at least from one segment to another). This failure often yields poor share predictions. The most popular remedy for this problem has been to use a mixture model. In this paper, we provide insights into this problem and illustrate the solution posed by latent class (LC) finite mixture models. We also describe several recent advances in the development of LC models for choice which have been implemented in a new computer program (Vermunt and Magidson, 2003a).

We conclude with a comparison of the LC finite mixture approach with the Hierarchical Bayes (HB) *continuous* mixture approach to choice modeling in a case study involving boots. We find that while both models provide comparable predictions, the LC models take much less time to estimate. In addition, the discrete nature of the LC model makes it more useful for identifying market segments and providing within-segment share predictions.

Introduction

The basic aggregate model as introduced by McFadden (1974) postulates that a choice of one alternative A_j is made from a set of alternatives A = $\{A_1, A_2, \ldots, A_J\}$, according to a random utility model

$U_j = V_j + e_j$

where V_j represents the systematic component of the utility and e denotes a stochastic error. The alternative A_j selected, is the one with highest utility U_j

In the simplest situation, the systematic utility component is assumed to satisfy a linear function of the choice attributes $X_1, X_2, \ldots, X_K$

$$V_j = b_{0j} + b_1 X_{j1} + b_2 X_{j2} + \ldots + b_K X_{jK}$$

$b_k X_{jk}$ is called the partworth utility associated with attribute k.

b_{0j} is called an alternative-specific constant, and may be omitted from the model.

Let Z denote the union of all the sets of alternatives. Then, under the assumption that e follows an 'Extreme Value Type 1 (Gumbel)' distribution, it follows that for any subset of alternatives $A' \subseteq Z$, the probability of choosing $A_j \in A'$ is given by the multinomial equation,

$$P_j = \exp(V_j) / \sum_{k \in A'} \exp(V_k) \quad ,$$

providing a probabilistic justification for this *conditional logit* model.

The implication of this equation is that if any alternative is excluded from the choice set A, its choice probability is allocated among the remaining alternatives proportional to their original choice probabilities. That is, it is assumed that none of the remaining alternatives is more likely (than any other) to serve as a substitute for the omitted alternative. Generally, this proportional-substitution-of-alternatives assumption, also known as IIA (Independence of Irrelevant Alternatives), does not hold in practice. McFadden (1974) recognized this as a weakness in his proposed model:

"This points out a weakness in the model that one cannot postulate a pattern of differential substitutability and complementarity between alternatives. … The primary limitation of the model is that the IIA axiom is implausible for alternative sets containing choices that are close substitutes."

The Latent Class Solution to the IIA Problem

LC Modeling assumes that IIA holds true within each of T ≥ 1 latent classes or segments:

$$P_{j.t} = \exp(V_{j.t}) / \sum_{k \in A'} \exp(V_{k.t}) \quad t = 1,2,\ldots,T$$

To illustrate the problem that occurs when the IIA assumption is violated and how the LC specification resolves this problem, consider the classic Red Bus/Blue Bus problem where the following 3 alternative means of transportation are available:

A_1 = Car, A_2 = Red Bus, A_3 = Blue Bus

For simplicity, we assume

$\exp(V_1) = .50$ and $\exp(V_2) = \exp(V_3) = .25$

In a choice between only A'_1 = Car and A'_2 = Red Bus, the aggregate model allocates the .25 Blue Bus probability between the remaining choices to preserve the 2:1 A_1: A_2 ratio of probabilities. This yields:

P(Car) ≡ P(A'_1) = .50/(.50+.25) = .67 which is clearly unreasonable.

In reality, the Red Bus would serve as a substitute for the Blue Bus yielding, P(A'_1) =.5.

With real data, LC modeling would reject the aggregate (1-class) model because it would not yield predicted choices between the Car and Red Bus (in a 2-alternative choice set) that are consistent with the *observed* choices. That is, the LC statistical criteria (discussed later) would reject the aggregate model in favor of T > 1 segments.

For simplicity, we will suppose that in reality there are 2 equal sized latent classes: those who prefer to take the bus (t=1) and those who prefer to drive (t =2).

Table 1

	exp(Vj)		
	Alternative j		
	1	2	3
Segment t	CAR	Red Bus	Blue Bus
1	0.02	0.49	0.49
2	0.98	0.01	0.01
Overall	0.50	0.25	0.25

In this case, the 2-Class LC model provides a more reasonable result. In the case the blue bus is no longer available, proportional allocation of its share over the 2 remaining alternatives separately within class yields:

P(Car.1) = .02/(.02 + .49) =.04,
P(Car.2) = .98/(.98 + .01) = .99

and overall, P(Car) = .5(.04) +.5(.99) = .52

The Red Bus/ Blue Bus problem illustrates the extreme case where there is *perfect* substitution between 2 alternatives. In practice, one alternative will not likely be a perfect substitute for another, but will be a more likely substitute than some others. Accounting for heterogeneity of preferences associated with different market segments will improve share predictions.

Latent Class Choice Modeling

Thus far we have shown that LC choice models provide a vehicle for accounting for the fact that different segments of the population have different needs and values and thus may exhibit different choice preferences. Since it is not known apriori which respondents belong to which segments, by treating the underlying segments as hidden or *latent* classes, LC modeling provides a solution to the problem of unobserved heterogeneity. Simultaneously, LC choice modeling a) determines the number of (latent) segments and the size of each segment, and b) estimates a separate set of utility parameters for each segment. In addition to *overall* market share projections associated with various scenarios, output from LC modeling also provides separate share predictions for each latent segment in choices involving any subset of alternatives.

Recent advances in LC methodology have resolved earlier difficulties (see Sawtooth Software, 2000) in the use of LC models associated with speed of estimation, algorithmic convergence, and the prevalence of local solutions. It should be noted that despite those early difficulties, the paper still concluded with a recommendation for its use:

> *"Although we think it is important to describe the difficulties presented by LCLASS, we think it is the best way currently available to find market segments with CBC-generated choice data"*

Advances in LC modeling

Several recent advances in LC choice modeling have occurred which have been implemented in a computer program called Latent GOLD Choice (Vermunt and Magidson, 2003a). These advances include the following:

- Under the general framework of LC regression modeling, a unified maximum likelihood methodology has been developed that applies to a wide variety of dependent variable scale types. These include choice, ranking (full, partial, best/worst), rating, yes/no (special case of rating or choice), constant sum (special case of choice with replication weights), and joint choices (special case of choice).
- Inclusion of covariates to better understand segments in terms of demographics and other external variables, and to help classify new cases into the appropriate segment.
- Improved estimation algorithm substantially increases speed of estimation. A hybrid a algorithm switches from an enhanced EM to the Newton Raphson algorithm when close to convergence.
- Bootstrap p-value – Overcomes data sparseness problem. Can be used to confirm that the aggregate model does not fit the data and if the power of the design is sufficient, that the number of segments in the final model is adequate.
- Automated smart random start set generation – Convenient way to reduce the likelihood of local solutions.
- Imposition of zero, equality, and monotonicity restrictions on parameters to improve the efficiency of the parameter estimates.
- Use of Bayes constants – Eliminates boundary solutions and speeds convergence.
- Rescaled parameters and new graphical displays to more easily interpret results and segment differences.

- New generalized R-squared statistic for use with any multinomial logit LC regression model.
- Availability of individual HB-like coefficients.

Each of these areas is discussed in detail in Vermunt and Magidson (2003a). In the next section we will illustrate these advances using a simple brand pricing example involving 3 latent classes.

LC Brand Pricing Example:

This example consists of six 3-alternative choice sets where each set poses a choice between alternative #1: a new brand -- Brand A (at a certain price), alternative #2: the current brand -- Brand B (at a certain price) and alternative #3: a None option. In total there are 7 different alternatives.

Table 2

Alternative	Brand	Price
A1	A	Low
A2	A	Medium
A3	A	High
B1	B	Low
B2	B	Medium
B3	B	High
None	None	

The six sets are numbered 1,2,3,7,8 and 9 as follows:

Table 3

	PRICE BRAND B		
PRICE BRAND A	Low	Medium	High
Low	1	4	7
Medium	2	5	8
High	3	6	9

Shaded cells refer to *inactive* sets for which share estimates will also be obtained (along with the six *active* sets) following model estimation.

Response data was generated[1] to reflect 3 market segments of equal size (500 cases for each segment) that differ on brand loyalty, price sensitivity and income. One segment has higher income and tends to be loyal to the existing brand B, a second segment has lower income and is not loyal but chooses solely on the basis of price, and a 3rd segment is somewhere between these two.

Table 4

	UpperMid	Loyal to Brand B	Price Sensitives
INCOME			
lower	0.05	0.05	0.90
lower middle	0.05	0.05	0.90
upper middle	0.88	0.10	0.02
higher	0.15	0.75	0.10

[1] The data set was constructed by John Wurst of SDR.

LC choice models specifying between 1 and 4 classes were estimated with INCOME being used as an active covariate. Three attributes were included in the models – 1) BRAND (A vs. B), 2) PRICE (treated as a nominal variable), and 3) NONE (a dummy variable where 1=None option selected). The effect of PRICE was restricted to be monotonic decreasing. The results of these models are given below.

Table 5

With Income as an active covariate:

LC Segments	LL	BIC(LL)	Npar	$R^2(0)$	R^2	Hit Rate	bootstrap p-value*	std. error
1-Class Choice	-7956.0	15940.7	4	0.115	0.040	51.6%	0.00	0.0000
2-Class Choice	-7252.5	14584.4	11	0.279	0.218	63.3%	0.00	0.0000
3-Class Choice	**-7154.2**	**14445.3**	**19**	**0.287**	**0.227**	63.5%	**0.39**	**0.0488**
4-Class Choice	-7145.1	14484.8	27	0.298	0.239	64.1%	0.37	0.0483

* based on 100 samples

The 3-class solution emerges correctly as best according to the BIC statistic (lowest value). Notice that the hit rate increases from 51.6% to 63.5% as the number of classes is increased from 1 to 3 and the corresponding increase in the $R^2(0)$ statistic[2] is from .115 to .287. The bootstrap p-value shows that the aggregate model as well as the 2-class model fails to provide an adequate fit to the data.

Using the Latent GOLD Choice program to estimate these models under the technical defaults (including 10 sets of random starting values for the parameter estimates), the program converged rapidly for all 4 models. The time to estimate these models is given below:

Table 6

	Time* (# seconds) to:	
LC Segments	Fit model	Bootstrap p-value
1	3	14
2	5	18
3	7	67
4	11	115

* Models fit using a Pentium III computer running at 650Mhz

[2] The R^2 statistic represents the percentage of choice variation (computed relative to the baseline model containing only the alternative-specific constants) that is explained by the model. In this application, the effect of the alternative-specific constants is confounded with the brand and None effects, and thus we measure predictive performance instead relative to the *null* model which assigns equal choice probabilities to each of the 3 alternatives within a set. This latter R^2 statistic is denoted by $R^2(0)$.

The parameters of the model include the size and part-worth utilities in each class. For the 3-class model they are given below.

Table 7

	Upper Mid	Loyal to B	Price Sensitives	
Size	0.35	0.33	0.32	Overall
$R^2(0)$	0.054	0.544	0.206	0.287

Attributes	Upper Mid	Loyal to B	Price Sensitives	p-value	Mean	Std.Dev.
BRAND						
A	-0.29	-1.15	0.03	2.1E-84	-0.47	0.50
B	0.29	1.15	-0.03		0.47	0.50
PRICE						
low	0.42	0.01	1.25	1.0E-53	0.55	0.51
medium	0.02	0.01	0.05		0.03	0.02
high	-0.44	-0.02	-1.30		-0.57	0.53
NOBUY	0.02	-1.04	0.62	1.4E-44	-0.14	0.68

The p-value tests the null hypothesis that the corresponding part-worth utility estimate is zero in each segment. This hypothesis is rejected ($p<.05$) showing that the effects are all significant. Notice that several within-segment utility estimates are close to zero[3]. In particular, the PRICE effect for the Loyal segment is zero[4] except for sampling variability. For this segment, the *unrestricted* PRICE effects turned out to be .00, .02, -.02 for the low, medium and high price levels respectively. The difference between the .00 and .02 reflect sampling error and are smoothed by the monotonicity restriction.

Viewing the part-worth utilities as random effects (see e.g. Vermunt and van Dijk, 2001, Vermunt and Magidson, 2002, 2003b,), we see that the brand A effect of -.29 (for the Upper Mid segment) occurs with overall probability .35, -1.15 with overall probability .33 and .03 with overall probability .32. Hence, the HB-like mean and Std Dev. Parameters can be computed as above. Similarly, individual HB-like part-worth parameters can be computed for each respondent, using that individual's posterior probability of being in each segment as weights in place of the overall probabilities.

The parameters in the model for predicting class membership as a function of INCOME are estimated simultaneously with the parameters given above. These include an intercept for the classes plus the direct relationship between INCOME and class membership. These estimates are expressed using effects coding in the following table. Identifying those estimates that are large in absolute value we see that class 1 (relative to the other classes) is more likely to have upper middle income, class 2 higher income, and class 3 lower and lower middle income. The low p-value shows that the relationship between INCOME and class membership is highly significant.

[3] Standard errors for these parameter estimates confirm that they do not differ significantly from zero.
[4] The PRICE effect for the Loyal to B segment could be further restricted to zero along with the BRAND effect for the Price Sensitive segment but for expediency this is not done in this paper.

Table 8

Model for Classes	Upper Mid	Loyal to B	Price Sensitives	
Intercept	-0.04	-0.28	0.31	
Covariates	Class1	Class2	Class3	p-value
INCOME				
lower	-0.56	-0.94	1.50	3.1E-88
lower middle	-0.97	-0.49	1.46	
upper middle	1.79	-0.13	-1.66	
higher	-0.26	1.56	-1.30	

Let us now return to the utility parameters. These log-linear parameters can be re-expressed in a form that allows easier interpretation, and that can also be used to develop an informative graphical display. To transform the parameters to column percentages, we can use a simple formula. For the BRAND effect parameter associated with brand A in segment t (betaA.t), we transform it into a column percentage form, ProbA.t, as follows:

ProbA.t = exp(betaA.t)/[exp(betaA.t)+exp(betaB.t)]

Thus, we obtain ProbA.t + ProbB.t = 1 for t=1,2,3

Table 9

	UpperMid	Loyal to B	Price Sens.
BRAND			
A	0.3605	0.0907	0.5155
B	0.6395	0.9093	0.4845

The interpretation of these numbers is as follows: In a choice between Brand A and Brand B where the other attribute (PRICE) is set at a common level, these re-scaled parameters represent the probability of choosing each Brand given latent class t. This interpretation holds regardless of the price level. Thus, for example the Price Sensitives are indifferent in their choice between brands A and B (.50 vs. .50) except for the nonsignificant difference in utilities (.03 vs -.03).

For attributes like price, these quantities indicate price sensitivities

Table 10

PRICE.A			
low	0.4769	0.3360	0.7237
medium	0.3198	0.3360	0.2192
high	0.2033	0.3281	0.0570

Using the overall latent class probabilities, these numbers can then be transformed to row %s to yield an insightful display.

Table 11
PROBMEANS Output -- Row %s

	Upper Mid	Loyal to B	Price Sensitives
Overall Probability	0.35	0.33	0.32
Attributes			
BRAND			
A	0.39	0.09	0.52
B	0.33	0.44	0.23
PRICE			
low	0.33	0.22	0.46
medium	0.38	0.38	0.24
high	0.36	0.55	0.09
NOBUY			
0	0.33	0.46	0.21
1	0.37	0.18	0.44
Covariates			
INCOME			
lower	0.08	0.04	0.88
lower middle	0.05	0.07	0.88
upper middle	0.86	0.10	0.04
higher	0.16	0.76	0.08

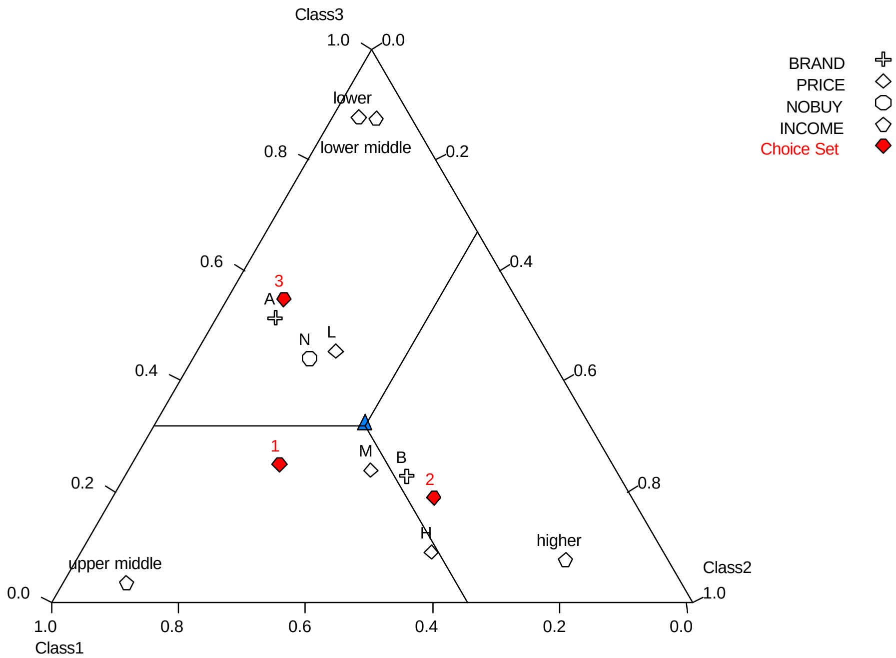
The same type of transformation from column to row percentages can be applied to the predicted choice probabilities.

Table 12

Column Percentages					Row Percentages			
			Price					Price
Set 1 n = 1350	Upper Mid	Loyal to B	Sensitives		Set 1	Upper Mid	Loyal to B	Sensitives
Choice 1	0.27	0.08	0.41		Choice 1	0.38	0.11	0.52
2	0.48	0.83	0.38		2	0.30	0.48	0.22
3	0.24	0.09	0.21		3	0.46	0.16	0.37
Set 2 n = 1350					Set 2			
Choice 1	0.20	0.08	0.17		Choice 1	0.46	0.18	0.36
2	0.53	0.83	0.53		2	0.29	0.43	0.27
3	0.27	0.09	0.29		3	0.43	0.14	0.43
Set 3 n = 1350					Set 3			
Choice 1	0.14	0.08	0.05		Choice 1	0.53	0.29	0.18
2	0.57	0.83	0.61		2	0.30	0.41	0.29
3	0.29	0.09	0.34		3	0.42	0.13	0.45
Set 4 n = 0					Set 4			
Choice 1	0.32	0.08	0.55		Choice 1	0.36	0.09	0.56
2	0.39	0.83	0.16		2	0.29	0.59	0.11
3	0.29	0.09	0.29		3	0.45	0.13	0.41
Set 5 n = 0					Set 5			
Choice 1	0.24	0.08	0.27		Choice 1	0.42	0.14	0.44
2	0.43	0.83	0.26		2	0.30	0.54	0.16
3	0.33	0.09	0.47		3	0.39	0.10	0.51
Set 6 n = 0					Set 6			
Choice 1	0.17	0.08	0.09		Choice 1	0.52	0.23	0.25
2	0.47	0.83	0.32		2	0.31	0.50	0.19
3	0.36	0.09	0.59		3	0.36	0.09	0.55
Set 7 n = 1350					Set 7			
Choice 1	0.38	0.08	0.63		Choice 1	0.36	0.08	0.56
2	0.29	0.82	0.05		2	0.26	0.70	0.04
3	0.34	0.09	0.33		3	0.47	0.12	0.41
Set 8 n = 1350					Set 8			
Choice 1	0.29	0.08	0.34		Choice 1	0.42	0.12	0.46
2	0.33	0.82	0.08		2	0.28	0.66	0.06
3	0.39	0.09	0.58		3	0.38	0.09	0.53
Set 9 n = 1350					Set 9			
Choice 1	0.21	0.08	0.12		Choice 1	0.53	0.20	0.28
2	0.36	0.82	0.11		2	0.29	0.63	0.08
3	0.43	0.09	0.77		3	0.35	0.07	0.58

Now, these row percentages can be used to position the corresponding attribute, covariate level, and choice on a common scale in a Barycentric coordinate display. For example, the following display plots the choices associated with set #6 (1:brand A Higher price vs. 2:brand B Medium price vs 3:None) together with INCOME and attribute levels in a common plot.

Figure 1: Barycentric Coordinate Display of 3-Segment Solution



In this plot, each segment corresponds to a vertex of the triangle. From this plot it can be seen that the lower right vertex corresponds to the Loyal to B segment. It is associated with Higher Income, Choice 2 (Brand B Higher Price) in set #3, brand B and Higher prices in general indicating the lack of price sensitivity. In contrast, the top vertex corresponds to the Price Sensitives. It is associated with Low and Lower Middle Income, a relatively higher preference for Brand A and choice 3 (None) when faced with the medium and higher priced set #3 options, Lower prices and the None option, more so than the other segments. Similarly, the lower left vertex corresponds to the UpperMid segment. It is associated with Upper Middle income and has a relatively higher likelihood of making choice #1 (Brand A Medium Price) in set #3.

Comparison between LC and HB

LC models utilize a discrete distribution of heterogeneity as opposed to a continuous distribution as assumed by HB. HB models assume that each respondent has their own unique preferences, while LC models assume that each respondent belongs to one of K (latent) segments, each of which has their own unique preferences. However, since the LC model also yields estimates for each respondent's probability of belonging to each segment, usage of these posterior probabilities as weights result in unique HB-like individual coefficients. Thus, even in the case that each individual has their own unique preference, an LC model containing a large number of classes can be used instead of HB to account for the heterogeneity. This approach avoids the necessity of making distributional assumptions (required by HB) which may cause poor predictions for cases with few responses (see e.g., Andrews et. al, 2002). This weighting is justified by viewing LC modeling as a non-parametric alternative to traditional HB-like random effects modeling (Vermunt and van Dijk, 2001).

To examine how LC and HB compare in practice, we used both in a CBC boot study (see Appendix).

Summary

When some alternatives in a set are more similar to each other than to the others, differential substitution is likely to occur which violates the IIA assumption in the aggregate model. In this case, use of the aggregate model is inappropriate and the resulting share predictions are distorted. The recommended remedy in this situation is to use a model that accounts for respondent heterogeneity of preferences. When comparing LC with HB models, LC modeling has the advantage of finding segments more directly than HB, and also is much faster to estimate. Moreover, a general maximum likelihood framework exists for LC models that include the ability to test various kinds of choice models, in restricted and unrestricted forms.

In our case study, prediction on hold-out sets (within-case/internal validation) of HB was found to be somewhat better than LC although there was a substantial fall-off in prediction error in the validation data, which is indicative of over-fitting. Our results appear to be consistent with Andrews et. al. (2002) who concluded:

> ”... models with continuous and discrete representations of heterogeneity … predict holdout choices about equally well *except* when the number of purchases per household is small, in which case the models with continuous representations perform very poorly”

Inclusion of covariates is a key issue to improve prediction on hold-out cases since the LC/HB models themselves don't improve prediction over the aggregate model in our case study. Although the application to hold-out cases is not emphasized in choice experiments we believe that it is an important topic. In our case study, the covariates were poor predictors of the LC segments so this aspect of the comparison was not addressed.

Our overall conclusion is that LC modeling has emerged as an important and valuable way to model respondent preferences in ratings-based conjoint and CBC, simulate choice shares and finding segments simultaneously. Several advances have been incorporated into a commercially available latent class tool called Latent GOLD Choice which substantially enhances both the speed and reliability of estimation. LC methods are now practical with traditional CBC data as well as with ranking (full, partial, best/worst), rating, yes/no (special case of rating or choice), and constant sum models, within choice sets. Compared to HB, LC is much faster, and directly provides insights regarding segments.

References:

Andrews, Rick L., Andrew Ainslie, and Imram S. Currim, 2002, "An empiracle comparison of logit choice models with discrete vs. continuous representations of heterogeneity", Journal of Marketing Research, Vol. XXXIX (November), 479-487.

McFadden (1974). Conditional logit analysis of qualitative choice behaviour. I. Zarembka (ed.), Frontiers in econometrics, 105-142. New York:Academic Press.

Natter, Martin and Markus Feurstein 2002, "Real world performance of choice-based conjoint models", European Journal of Operational Research, 137, 448-458.

Sawtooth Software (2000) "CBC Latent Class Analysis Technical Paper". Sawtooth Software Technical paper Series,2000.

Vermunt, Jeroen K. and Liesbet A. van Dijk. "A non-parametric random-coefficient approach: the latent class regression model", in Multilevel Modelling Newsletter, 2001, 13, 6-13.

Vermunt, J.K. and Magidson (2003a) "Latent GOLD Choice 3.0 Users Guide" Belmont, MA: Statistical Innovations.

Vermunt, J.K. and Magidson "Non-parametric Random Effects Model", in Encyclopedia of social science methodology, Sage, forthcoming 2003b.

Vermunt, J.K. and Magidson J., (2002), "Latent Class Models for Classification", Computational Statistics and Data Analysis.