A Note of Caution Regarding Applying Basic Latent Class Analysis

Ji Hee Song and Richard J. Fox

We investigate latent class analysis as a market segmentation method. In particular, we explore the ramifications of the often assumed, perhaps necessarily, local independence condition. Using survey data for a major soft drink brand, we identify brand characteristics which are perceived very differently across simply defined demographic segments. We next apply latent class analysis, assuming local independence, to these same characteristics to derive a segmentation structure. These derived segments, driven by the local independence assumption, are completely different from the initial demographic segments. Strong justification for this assumption, which dictates the solution and cannot be explored in advance, clearly should be present.

Keywords: Latent Class Analysis, Market Segmentation, Local Independence

Introduction

Market segmentation is a vital step in the process of developing a marketing strategy. The goal of market segmentation is to organize consumers into groups, similar within, and very different between, in terms of their preferences for products, usage of products, responses to marketing mix strategies, and so forth. Segmentation approaches can be forward (e.g., segmentation on the basis of consumer characteristics such as demographics), backward (e.g., on the basis of preferences or purchase behaviors), or hybrid, employing both forward and backward elements (Andrews & Currim 2003). Latent class analysis, often applied to observed behavior or attitudes as a backward approach, has become a fairly popular segmentation method in recent years. For example, Grover and Srinivasan (1987) applied this method to a brand-switching matrix to determine market structure and market segments. The derived segments differed in size and with respect to probabilities of choosing the different brands in the product class. Since they first introduced this method to market structuring and segmentation, latent class analysis has been employed in various marketing segmentation studies (Kamakura & Russell 1989, Zenor & Srivastava 1993).

Critical assumptions of the latent class approach adopted by Grover and Srinivasan (1987) were stationarity of the brand choice process, local independence, i.e., zero order brand choice process within segment, and homogeneity of brand choice probabilities within segment. In later research, these assumptions have been relaxed to accommodate marketing mix variables and more sophisticated models of consumer choice. For instance, Kamakura and Russell (1989) developed a latent class model based on a logit choice model, where the consumer segments differed in both brand preferences and price sensitivity.

Despite the fact that underlying choice models can be quite sophisticated, the notion of local independence, i.e., zero order choice process within segment, remains a key assumption in many applications, e.g., Grover and Srinivasan (1987). Further, this assumption simplifies the model, and hence reduces the number of parameters to be estimated, which may be crucial to making the latent class approach possible. Local independence means that measurement variables are statistically independent at the segment or "local" level, but not overall. For example, consecutive choices of brands in a particular category are dependent when viewed overall, but independent when viewed within segments. This assumption, which, as noted above, is crucial to the statistical model used in the analysis, cannot be investigated in advance because the segments are latent. Hence, when using this approach in practice, one must argue that this assumption is reasonable given the context. What are the consequences of applying latent class analysis based on this assumption?

The objective of this paper is to examine the consequences of adopting the latent class method, assuming local independence, in market segmentation analysis and to demonstrate how this assumption "drives" the solution. Again, practitioners may need to make this assumption, without being able to explore its validity, in order to apply the latent class approach to market segmentation. In our study, respondents to a marketing research survey, which focused on attitude and behavior toward a specific soft drink brand, are first segmented on the basis of simple demographic variables, gender and age (teen versus adult). Attitudes (toward a specific brand) are analyzed across the demographic segments to identify those brand characteristics which are clearly perceived differently among the segments. Hence, a set of attributes, for which large differences in brand perceptions across crisp and simple demographic segments are observed, are determined. It is then shown that responses to the questions corresponding to perceptions of the brand on these selected attributes clearly do not satisfy the local independence assumption within the demographic segments. The focus of the paper is to examine the degree to which this simple market structure, albeit an artificial one, is visible when the segmentation solution is derived by working "backwards" and applying latent class analysis, assuming local

independence, to the brand ratings data for the same set of attributes. The segmentation which emerges from the latent class analysis is very different from the initial demographic segmentation, and demonstrates the powerful influence of the local independence assumption on the solution.

In the following section, brief reviews of the latent class analysis method and the local independence assumption are provided. Next, we describe the data set used for the analyses and specific segmentation model. Results are presented in the subsequent section, and discussion and research implications are provided in the last section.

Latent Class Analysis and Local Independence Assumption

Latent class analysis has been used to study patterns of interrelationships among observed variables in order to develop underlying latent segments (Mccutcheon 1987). As noted earlier, Grover and Srinivasan (1987) used a latent class approach to derive market structure and latent segments from a brand-switching matrix based on two consecutive brand choices in the coffee category by members of commercial household panel. The latent class concept has been used with various model-building approaches (e.g., logit choice and regression models). For example, Kamakura and Russell (1989) used a latent class logit model to develop market segments that vary with respect to basic brand appeal and price sensitivity based on brand choices made by a commercial household panel. In the research reported in this paper, we assume the observed variables reflect simple choice data such as two consecutive brand choices, as considered by Grover and Srinivasan (1987), or categorical data (e.g., whether a person subscribes to specific magazines or not). In this context, the local independence premise is that the observed (manifest) variables are independent within segment, but not at the "global" level.

As noted earlier, the model simplicity resulting from the local independence assumption may be vital to the analysis. A necessary condition for identification in latent class analysis is that the degrees of freedom be positive (Mccutcheon 1987). The degrees of freedom is equal to the number of observed cells (response categories based on cross-classification of the levels of the observed variables) less one, minus the number of estimated model parameters (Press 1972). Local independence (all observed variables are independent within categories of the latent variable, i.e., segments) plays an important role in reducing the number of model parameters by simplifying the model, and thus satisfying the necessary condition for model identification. Therefore, it may be necessary to assume local independence in order to employ a latent class

approach to segmentation based on the data which are available. However, assuming local independence determines a particular structure for the observed data, and guides determination of a latent factor that explains all observed dependencies (Anderson 1988).

We may explain the principle of local independence using an example taken from Mccutcheon (1987). Suppose that a group of 300 persons is asked whether they have read that last issue of magazines A and B. Their responses are shown in Table 1 below. These data indicate that the two variables are quite strongly related to one another. Readers of A are far more likely to read B than non-readers of A. The value of the chi-square statistic for testing the hypothesis of independence of these two variables (read A, read B) is 8.42 which is highly significant ($p<.01$). In other words, readership of magazine A and readership of magazine B are related, i.e., not independent.

Table 1. Cross-Tabulation with Readership of Magazine A & B

	Read B	Did not read B	Total
Read A	95	55	150
Did not read A	70	80	150
Total	165	135	300

Chi-square= 8.42 ($p < .01$)

Now suppose information regarding the respondents' income levels (i.e., high vs. low) is available. The 300 people can be divided into two groups, as shown in Table 2. For each of the two income groups, readership of the two magazines (A and B) are independent. The reading behavior is, however, very different in the two groups: readership of both magazines is relatively high in the high income group, and low in the low income group. The association between the readership of magazine A and magazine B is explained by the third factor, income level.

Table 2. Cross-Tabulation with Income and Readership of Magazine A &B

	High Income (n=150)			Low Income (n=150)		
	Read B	Did not Read B	Total	Read B	Did not Read B	Total
Read A	80	20	100	15	35	50
Did not read A	40	10	50	30	70	100
Total	120	30	150	45	105	150

Chi-square= .047 (p= .83) Chi-square= .036 (p= .85)

The criterion of local independence provides a means for determining whether relationships among observed variables are due to some other (perhaps not even measured) variable(s). When a set of interrelated variables (readership of magazine A and B in the above example) are found to be locally independent within categories of some other variable (income level in our example), we say that the additional variable "explains" the observed relationships. Therefore, the local independence assumption is necessary to identify the underlying explanatory factors. Some researchers (e.g., Clogg 1988) consider this property an axiom rather than an assumption. Latent class analysis, assuming local independence, "forces" a solution (segments) that satisfies the assumption as well as possible, but may not be meaningful or valuable (as we later demonstrate) at all. Further, the latent explanatory factor(s) may be difficult to identify in light of available information for profiling the segments. Finally, the local independence assumption cannot be investigated a priori, so strong justification for this assumption should be present before using it in conjunction with applying the latent class approach.

Data

The data used for the study are based on a survey of over seven thousand consumers who were questioned regarding their attitudes and behavior toward a major soft drink brand. Respondents were asked to answer "Yes" or "No" to a sequence of questions asking whether the brand possessed each of 30 attributes. The attributes are shown in Table 3, along with symbols used for displaying them in the perceptual maps presented later.

Analysis

The sample was first partitioned into four demographic segments based on gender and age, teen versus adult (Table 4). Correspondence analysis (Hair, Anderson, Tatham, & Black 1998) was used to develop a perceptual map depicting associations between these four segments and the 30 attributes. Seven attributes which clearly separate the four segments were selected. The condition of local independence for all pairs of the seven attribute ratings was investigated within these four segments, and found not to apply at all. Then, using the latent class approach, assuming local independence, market segments were derived using the seven questions corresponding to the seven attributes which were found to differentiate the four segments, as the "manifest" observed variables. The "new" segments were then compared with the initial segments.

Table 3. Attribute Variables and Their Symbols

Variables	symbol	Variables	symbol
Very Refreshing	a	A Fun Soft Drink	p
Quenches My Thirst	b	Gives Me Enjoyment Anytime	q
Gives Me a little Extra Energy	c	Adds a Little Magic to My Life	r
Is Great Tasting	d	Is Part of my Daily Life	s
Has a Clean Taste	e	Drink I and Friends Like to Share	t
Goes Well With Food	f	Helps Me Bring Happiness Family	u
Good Times With Family Better	g	Cool	v
For Someone Like Me	h	Talks to Me in an Honest Way	w
Appropriate in Social Situations	i	Encourage Do Things Your Way	x
For When Having Fun with Friends	j	Says no Matter Consequences	y
Helps Feel Free Express Myself	k	Soft Drink Most Engaging Style	z
Worth What it Costs	l	Most Exclusive Soft Drink	aa
High Quality	m	For Palates Distinct Taste	ab
Recognized As Most Admired Brand	n	For People Know What They Like	ac
Rich Full Bodied Taste	o	The Most Natural Soft Drink	ad

Table 4. Initial Four Segments based on Demographic Variables

Segment	Characteristics		N	(%)
1	Male/Teenagers	(12-19 Years)	1206	(17.0%)
2	Female/Teenagers	(12-19 Years)	717	(10.1%)
3	Male/Adults	(20-49 Years)	2314	(32.7%)
4	Female/Adults	(20-49 Years)	2840	(40.1%)

Initial Market Structure

A simple approach to market segmentation is to organize consumers into groups based on demographic variables. Such segmentation schemes are typically very practical because marketing tactics, capitalizing on the peculiarities of the respective groups, are easily directed at these separate segments. The four crisp demographic segments defined by crossing age, teen versus adult, with gender (Table 4) are clearly highly actionable. Furthermore, if data which clearly discriminates these four groups from each other were used as the basis for developing a

segmentation scheme, it seems desirable that these four segments would emerge, to some degree, from the method of analysis. Correspondence analysis was used to identify the attributes which clearly discriminated among these four demographic segments, and the responses to the corresponding brand rating questions became the basis for the latent class segmentation.

Correspondence Analysis

To determine the attribute variables which clearly separate the four demographic segments, a preliminary correspondence analysis was performed on the frequency table containing the number of "YES" responses for each attribute (row) and segment (column). The resulting perceptual map is based on associations and disassociations, between segments and attributes, based on observed and expected frequencies within each cell, (Hair et al. 1998). One of the distinctive features of correspondence analysis is the ability to simultaneously graphically portray the relationships between the variables and/or segments (Bendixen 2003). Seven attributes which clearly "define" the four demographic subgroups were chosen based on this map. These seven attributes served as the manifest variables for the latent class analysis.

Latent Class Analysis

Each consumer is asked to respond yes (1) or no (0) regarding whether each of K=7 attributes describes the particular soft drink brand. A response therefore consists of a sequence of K=7 digits, each 0 or 1. The goal of the latent class model is to partition the sample into m mutually exclusive and exhaustive groups. Let ν_α denote the probability that an individual comes from the α th group, $\alpha = 1,\ldots,m$, and let $\lambda_{\alpha i}$ denote the probability that an individual from group α responds "yes" for variable *i*, *i* =1, 2,…,7. Assuming local independence, the probability of a particular sequences of 1's and 0's for a consumer in segment α is the product of the respective $\lambda_{\alpha i}'s$ and $(1-\lambda_{\alpha i})'s$. Let π_i denote the probability a randomly selected consumer responds "yes" for attribute *i*; π_{ij} denote the probability a randomly selected consumer responds "yes" for both variables *i* and *j*; π_{ijk} denote the probability of "yes" for variables *i*, *j*, and *k*; and so on. These probabilities are represented as follows (Press 1972):

$$\pi_i = \sum_{\alpha=1}^{m} \nu_\alpha \lambda_{\alpha i}$$

$$\pi_{ij} = \sum_{\alpha=1}^{m} \nu_\alpha \lambda_{\alpha i} \lambda_{\alpha j}$$

$$\pi_{ijk} = \sum_{\alpha=1}^{m} \nu_\alpha \lambda_{\alpha i} \lambda_{\alpha j} \lambda_{\alpha k}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\pi_{ijk\cdots z} = \sum_{\alpha=1}^{m} \nu_\alpha \lambda_{\alpha i} \lambda_{\alpha j} \lambda_{\alpha k} \cdots \lambda_{\alpha z}$$

$$where\ \nu_\alpha \geq 0,\ \alpha = 1, \cdots, m,$$

$$\sum_{\alpha=1}^{m} \nu_\alpha = 1,$$

$$and\ for\ \alpha = 1, \cdots, m\ and\ i = 1, \cdots, 7$$

$$0 \leq \lambda_{\alpha i} \leq 1.$$

The data used for the latent class analysis consist of the frequencies of respondents classified into the 2^7 = 128 cells based on their responses to the seven attribute questions. These observed sample frequencies provide estimates of $\pi_i, \pi_{ij}, \ldots,$ which are used to develop maximum likelihood estimates of v_α, $\alpha = 1, \cdots\cdots, m,$ and $\lambda_{\alpha i}\ i = 1, \cdots\cdots 7$, for a given number of segments, m. CDAS software (Eliason 1997) was used to perform the maximum likelihood estimation. The associated degrees of freedom, for a given m, are (2^7-1)-[(m-1)+7m], which is the number of cells less one, minus the number of estimated parameters ((m-1)segment probabilities plus 7(m) latent probabilities, i.e., 7 λ's, for each segment). As noted earlier, a necessary condition for identification is that the degrees of freedom be positive. Hence, in our case, a maximum of m=15 segments can be included in the model.

Typically, the fit of a model can be assessed by applying the Chi-Square Goodness-of-Fit test to the observed vs. expected frequencies, determined using the estimated parameters, for the 2^7 cells. However, the disadvantage of the Chi-Square statistic is that it is directly proportional to the sample size. Therefore, for large sample sizes, such as is the case in this application, the test is overly powerful. Following the approach used by Grover and Srinivasan (1987), we use an adjusted R-square as a goodness-of-fit measure in this paper. Adjusted R-square provides measures of improvement in fit obtained by segmentation and is calculated as:

$$\overline{R}^2 = \frac{\frac{\chi^2(1)}{d.f(1)} - \frac{\chi^2(m)}{d.f(m)}}{\frac{\chi^2(1)}{d.f(1)}}$$

where $d.f(m)$ is the degrees of freedom for the m segment model.

Results

Determining the Attributes

The first step in analyzing the results of correspondence analysis is to determine the number of dimensions required to reasonably portray the relationships in the data. There is usually a trade-off between the explanatory power and interpretability (dimensions). The dimensionality and the "percent of inhomogeneity" (i.e., percentage of the typical Chi-square measure of deviations of observed frequencies from expected frequencies) explained are shown in Table 5. As can be seen in the table, 95.99% of "inhomogeneity" can be explained by a two-dimensional solution. Therefore, a two-dimensional solution is more than adequate to represent the associations and disassociations among the initial four segments and the attributes.

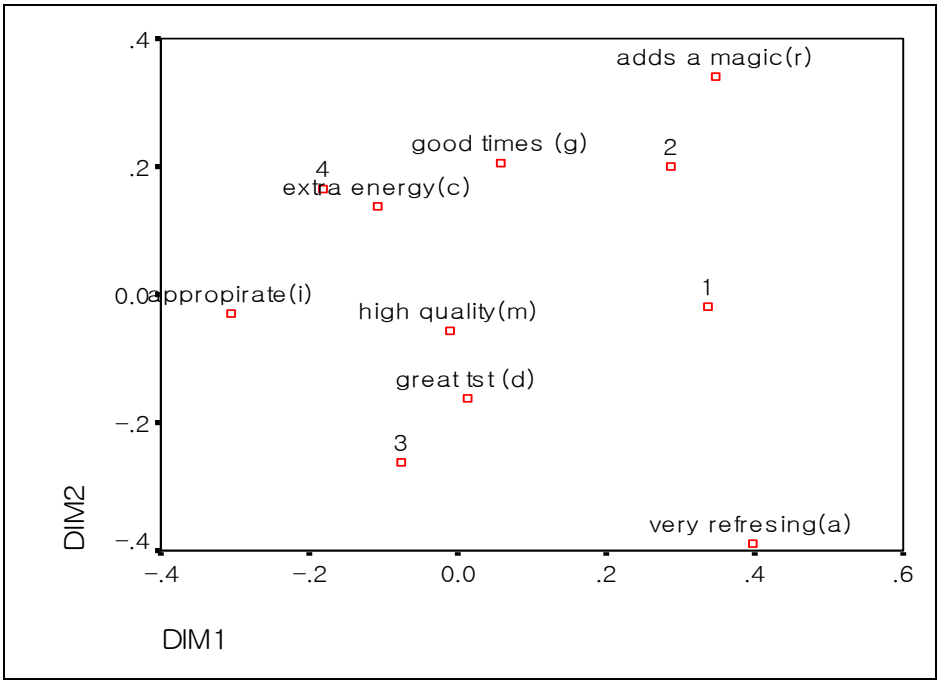
Table 5. Dimensionality and "Percent of Inhomogeneity" Explained

Dimension	Proportion Explained %	Cumulative Explained %
1	65.2	65.2
2	30.8	96.0
3	4.0	100.0
Total	100.0	

The perceptual map displaying the associations/disassociations between segments and attributes was analyzed and variables that clearly distinguish the four demographic segments from each other were determined. Seven attributes were identified as characteristics for which perceptions differed considerably among the segments. To present a clear picture, another correspondence analysis was conducted using only these seven selected variables and four segments. Again, a two-dimensional solution captures the associations/disassociations

very well (98.16% of inhomogeneity is explained by a two-dimensional solution). The resultant map is shown in Figure 1. (Table 3 shows the letter codes for the attributes, and the segments are represented numerically according to Table 4.)

Figure 1. Perceptual Mapping of Market Segments and Seven Attributes



As can be seen from Figure 1, "very refreshing" is strongly associated with Segment 1 (teenage males), and strongly disassociated with Segment 4 (female adults). "Adds a little magic to my life" was strongly associated with Segment 2 (teenage females), and disassociated with Segment 3 (male adults). Segment 4 (female adults) responded that the brand "gives me a little extra energy" more than expected, but Segment 1 (teenage males) responded in this way less often than expected. Similar observations can be made for the other attributes included among the seven analyzed.

Violation of Local Independence Assumption

The validity of the local independence assumption for the 7 brand attribute questions and the initial demographic segmentation was examined. The tests for independence between two attribute variables (Chi-Square test) are significant ($p<.01$) for all pairs of attribute variables within each of the four demographic segments, clearly indicating that the attribute variables are dependent within each segment. The local independence assumption is clearly violated in

the initial demographic segmentation. What happens if one makes this assumption and applies latent class analysis to develop a segmentation model? How far from this demographic segmentation is the solution "forced" by the assumption of local independence?

Latent Structure Segmentation/Determining the Number of Segments

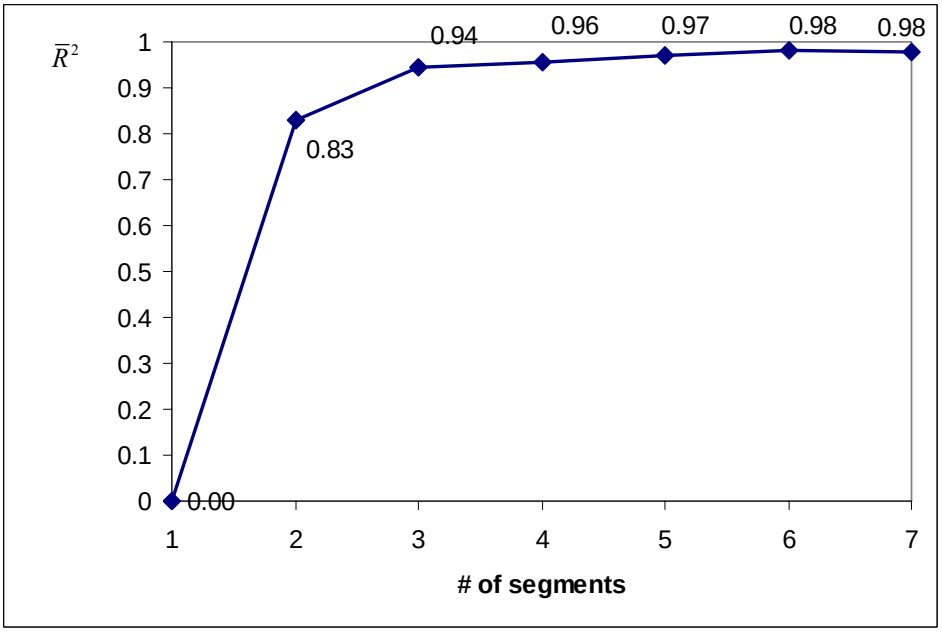
The typical latent class analysis segmentation was performed varying the number of segments from one to seven (the maximum number of segments satisfying the necessary condition for model identification is 15). The corresponding degrees of freedom, Chi-square goodness-of-fit statistic, p-value, and adjusted R-square are given in Table 6.

Table 6. DF, Chi-Square, P-value and $\overline{R}^2$ of the Model

Segments	DF	Chi-square	p-value	Adjusted R-square
1	120	9134.64	0.00000	0
2	112	1458	0.00000	0.828987
3	104	439	0.00000	0.944548
4	96	328	0.00000	0.955116
5	88	209	0.00000	0.9688
6	80	124	0.00118	0.979638
7	72	117	0.00064	0.978653

The p-value is small enough to reject the null hypothesis of appropriateness of the model for any number of segments, 1 thru 7. However, considering the disadvantage that the Chi-square measure tends to be significant for very large sample sizes such as ours, we expect the Chi-square test to reject, i.e., have a low p-value, and we examine the adjusted R-square values. Figure 2 shows the value of adjusted R-square for the number of segments. The adjusted R-square improves up to m=6, and stays flat thereafter. However, there is only a marginal increase after m=4 ($\Delta\overline{R}^2$ = .01). Hence, in the interest of parsimony, and convenience of comparisonwith the four demographic segments, the four-segment solution was adopted. The adjusted R-square corresponding to m=4 is .96, which indicates very good-fit. (Also, the parameters are known to be estimable when m=4 as noted by Press (1972, p. 321), i.e., m≤4 is a sufficient condition for identification in this case.)

Figure 2. Goodness-Of-Fit $\overline{R}^2$ For Different Numbers of Segments



Four-Segment Solution

The derived four segments are approximately equal in size; the segment probabilities were estimated to be 28%, 26%, 21%, and 25% (prior membership probabilities, v_α's). Posterior probabilities of segment membership were calculated for each of the 128 cells of respondents using the estimated model parameters, the respective sequences of 7 choices defining the cell, and the prior probabilities of segment membership. Each cell was assigned to the most "likely" segment based on posterior probabilities of membership. As is typical, these "pseudo segments" were treated as surrogates for the latent segments in order to gain understanding of the compositions of the latent segments. The four pseudo segments, labeled A, B, C and D, contain 30%, 27%, 21%, and 23% respectively of the respondents (see Table 7), and these percentages agree closely with the estimated sizes of the segments (v_α's).

Table 7 shows the age (teenager versus adult) and gender composition of the four pseudo segments. These segments do not relate to the initial four demographic segments at all. Rather, there are consistent patterns for gender and age within each of the four pseudo segments. Every pseudo segment is about evenly divided between males and females, and there is little variation in age composition. The segments derived thru latent class analysis are

drastically different from the initial demographic segments, and exhibit no difference in age and gender, detracting from any operational value.

Table 7. Demographic Profiles and Sizes of Derived Segments

	Derived Segment				
	A	B	C	D	
Gender					
% Male	53	50	49	46	
% Female	47	50	51	54	
Age					
% Teenager	28	32	19	29	
% Adult	72	68	81	71	
Size					
(# of Respondents*)	2089	1981	1466	1631	7077
	(30%)	(27%)	(21%)	(23%)	

*Based on classifying each of the 128 cells of respondents into the derived segment having the highest posterior probability of membership. Posterior probabilities are determined from prior probabilities and responses to the seven attribute questions.

To examine how the two segmentations are related, a cross-tabulation of respondents with initial demographic segments as rows and latent class segments (actually pseudo segments) as columns was constructed. As shown in Table 8, each initial segment is well spread over the surrogate latent class segments, indicating again that the latent class model does not come at all close to producing the pre-existing market structure based on demographics.

As noted earlier, local independence was assumed (as is typical) in the latent class analysis. The cross-tabulation analysis of responses to pairs of attribute questions within the pseudo segments derived via latent class analysis reveals that most Chi-square statistics are not significant at the 10% level of significance, indicating that the local independence assumption is generally satisfied within these pseudo segments. As expected, the latent class analysis produces solutions which meet the local independence assumption. However, one must argue conceptually to justify making this assumption. Taking this assumption for granted, without much consideration, dictates a unique type of solution, which may not be very useful.

As seen in this example, the method can lead to a segmentation, far removed from the simple, crisp and highly actionable existing market structure based on demographic variables.

Table 8. Cross-Tabulation with Initial Segments and Latent Class Segments

		Latent Segment				Total
		Seg A	**Seg B**	**Seg C**	**Seg D**	
Initial	Seg 1	373	373	185	275	1206
Segment	Seg 2	203	232	91	191	717
	Seg 3	733	571	539	471	2314
	Seg 4	780	715	651	694	2840
Total		2089	1891	1466	1631	7077

Each of the surrogate latent class segments is described by responses to the 7 attribute questions of its estimated survey composition (pseudo segments) in Table 9. Segment B is the strongest for the brand, and the brand scores very high within the segment for all the attributes. On the other hand, Segment C generally has a very low opinion of the brand, and the brand scores relatively low on all 7 attributes. It is important to point out again that these derived segments cannot be differentiated from each other using the two demographic variables age and gender. There are basically no differences among these segments with respect to these demographic variables. Thus, targeting these groups with custom designed marketing efforts appears difficult at best.

Table 9. Profiles of Derived Segments Across Seven Attributes

	Seg A	Seg B	Seg C	Seg D
	30%	27%	21%	23%
	Within-segment proportions			
	%	%	%	%
Appropriate in Social Situations	80.1	95.5	27.1	74.7
Gives Me a little Extra Energy	25.5	70.4	2.2	21.8
Good Times With Family Better	55.3	98.7	0.0	55.9
High Quality	90.4	97.7	11.7	59.5
Is Great Tasting	100.0	93.5	3.6	12.2
Adds a Little Magic to My Life	8.4	87.8	0.0	23.9
Very Refreshing	27.9	61.8	2.4	15.5

In contrast, Table 10 is similar to Table 9, but shows percentages calculated for the initial four demographic segments. None of the profiles of the initial four demographic segments (Table 10) match any of the four surrogate latent class demographic segments (Table 9). It is important to note that the correspondence analysis map shown in Figure 1 is based on the same frequency data used to produce Table 10. However, correspondence analysis adjusts for the expected frequencies in each cell as determined by the respective row and column marginals. Simply examining raw percentages, such as those in Table 9 can be deceiving

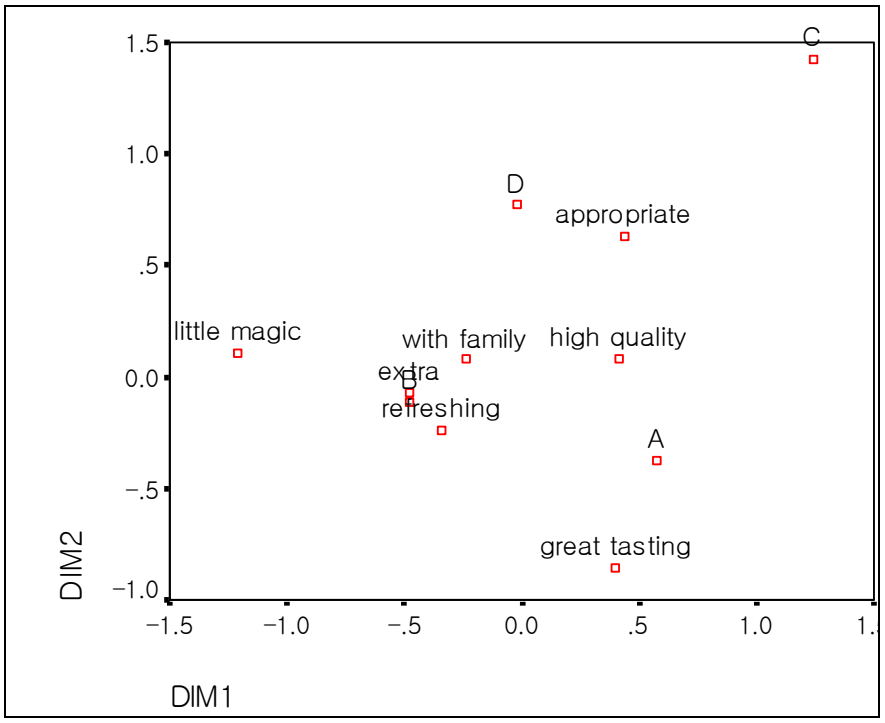
Table 10. Profiles of Initial Segments Across Seven attributes

	Seg 1	Seg 2	Seg 3	Seg4
	17.0%	10.1%	32.7%	40.1%
	Within-segment proportions			
	%	%	%	%
Appropriate in Social Situations	69.4	73.9	71.8	72.8
Gives Me a little Extra Energy	33.9	34.9	30.0	32.4
Good Times With Family Better	60.9	66.1	50.6	54.8
High Quality	75.7	73.8	67.2	66.2
Is Great Tasting	64.3	61.5	58.1	54.5
Adds a Little Magic to My Life	38.0	40.7	26.8	30.1
Very Refreshing	35.2	33.9	29.9	23.9

when looking for associations/disassociations between segments and specific attributes. For example, Segment B of Table 9 exhibits a relatively high level of agreement on all seven attributes, and an "adjustment" according to expected frequencies, as is done in correspondence analysis, should be made to understand which attributes are "really" associated with the brand within this segment.

To understand which attributes are associated/disassociated with which latent class segments, a correspondence analysis map, based on Table 9 and shown in Figure 3, was developed. The relationships are quite different from those shown in Figure 1, which is based on the results of correspondence analysis using the initial demographic segments. For example, notice how much more influential "great tasting" and "appropriate for social occasions" are in separating the segments in Figure 3 relative to Figure 1. The reverse is true for "very refreshing". In summary,
the segments derived from the latent class model do not agree, even marginally, to the initial segments defined by demographic variables on the context of brand perceptions in the selected attributes. The latent class segments are not related to the two demographic variables and exhibit different patterns of brand perceptions across the attributes.

Figure 3. Correspondence Analysis Map of "Derived" Segments and Seven Attributes



Discussion

In this paper, we investigated latent class analysis as a market segmentation method. In particular, we explored the ramifications of the local independence assumption typically assumed in this approach. Using survey data regarding consumer perceptions of a major soft drink brand, we identified four distinct segments based on simple typical demographic characteristics. Using correspondence analysis to graphically portray relationships among the attributes and the demographic segments, we determined seven brand attribute ratings that clearly differentiate these four segments from each other. Given this market structure, we applied latent class analysis, assuming local independence, using these same seven attributes to derive a segmentation structure. The latent class analysis approach yielded four segments of about equal size. Posterior probabilities of segment membership were used to assign respondents to segments, and thus to construct four pseudo or surrogate latent segments to represent the segmentation derived using latent class analysis. We compared the surrogate segments with the initial segments, and observed that, not unexpectedly, the two segmentation structures are drastically different. The initial demographic segmentation does not satisfy the local independence assumption, while, as expected, this condition is satisfied by the surrogate segmentation derived using the latent structure approach. The latent class solution is driven by the local independence assumption, which basically overwhelms the intrinsic associations and disassociations between demographic segments and attributes imposed on the data. The resulting solution is peculiar in this way, and, in this example, provides a solution that deviates dramatically from a useful and simple existing market structure.

Latent class analysis has been used frequently in market structure/segmentation applications. However, very little research has examined the consequences of the crucial local independence assumption often made by necessity when using this method. Our findings show that latent class analysis, driven by the local independence assumption, may suggest a market structure which is not very useful, and is far removed from a simple and useful segmentation scheme which does not satisfy the local independence assumption. Therefore, a simple straightforward approach might be more productive than a more sophisticated approach such as latent class analysis, which is data-driven and highly dependent on an assumption which cannot be investigated in advance. Researchers and practitioners need to be cautious in applying this approach for market segmentation. Strong justification for the local independence assumption should be present before proceeding this way.

References

Anderson EB (1988). Comparison of Latent Structure Models in R Langeheine and J Rost (eds). *Latent Trait and Latent Class Models*. New York: Plenum.

Andrews RL & Currim IS (2003). Recovering and profiling the true segmentation structure in markets: an empirical investigation. *International Journal of Research in Marketing 20*, 177-192.

Bendixen M (2003). A practical guide to the use of correspondence analysis in marketing research. *Marketing Bulletin*, *14*.

Clogg CC (1988). Latent Class Models for Measuring. in R Langeheine and J Rost (eds). *Latent Trait and Latent Class Models*. New York: Plenum.

Eliason SR (1990). *The Categorical Data Analysis System Version 3.50 User's Manual*. University of Iowa.

Eliason SR (1997). *The Categorical Data Analysis System Supplemental User's Manual for Command Line*. University of Iowa.

Grover R & Srinivasan V (1987). A simultaneous approach to market segmentation and market structuring. *Journal of Marketing Research*, *24 (May)*, 139-153.

Hair JF, Anderson RE, Tatham RL & Black WC (1998). *Multivariate Data Analysis.* New Jersey: Prentice Hall.

Kamakura WA & Russell GJ (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, *26 (November)*, 379-390.

Mccutcheon AL (1987). *Latent Class Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills: Sage.

Press SJ (1972). Applied Multivariate Analysis. New York: Holt, Rinehart and Winston.

Zenor MJ & Srivastava RK (1993).Inferring market structure with aggregate data: A latent segment logit approach. *Journal of Marketing Research*, *25 (August)*, 369-379.

Ji Hee Song is a doctoral candidate and Richard J. Fox is an Associate Professor in the Department of Marketing, Terry College of Business, University of Georgia