27/01/19

LATENT CLASS MODELS

Jay Magidson,
Statistical Innovations Inc.

Jeroen K. Vermunt,
Tilburg University

This is an updated version of Magidson, J., and Vermunt, J.K, (2004). Latent class models. D. Kaplan (ed.), The Sage Handbook of Quantitative Methodology for the Social Sciences, Chapter 10, 175-198. Thousand Oaks: Sage Publications.

1. INTRODUCTION

Latent class (LC) modeling was initially introduced by Lazarsfeld and Henry (1968) as a way of formulating latent attitudinal variables from dichotomous survey items. In contrast to factor analysis, which posits continuous latent variables, LC models assume that the latent variable is categorical, and areas of application are more wide-ranging. The methodology was formalized and extended to categorical variables with more than two categories by Goodman (1974a, 1974b) who also developed the maximum likelihood (ML) algorithm that serves as the basis for many of today's LC software programs. In recent years, LC models have been extended to include observable variables of mixed scale type (nominal, ordinal, continuous and counts), covariates, and to deal with sparse data, boundary solutions, and other problem areas.

In this chapter, we describe three important special cases of LC models for applications in cluster, factor and regression analysis. We begin by introducing the LC cluster model as applied to nominal variables (the traditional LC model), discuss some limitations of this model and show how recent extensions can be used to overcome them. We then turn to a formal treatment of the LC factor model and an extensive introduction to LC regression models before returning to show how the LC cluster model as applied to continuous variables can be used to improve upon the K-means approach to cluster analysis. We use the Latent GOLD computer program (Vermunt and Magidson, 2000, 2015) to illustrate the use of these models as applied to several data sets.

2. TRADITIONAL LATENT CLASS MODELING

Traditional LC analysis defines a model for a set of categorical (or nominal) variables. Suppose we are dealing with 4 nominal observed (manifest) variables, which we denote by $y_1,\ y_2,\ y_3,$ and y_4. The number of categories of these variables is denoted by M_1, M_2, M_3, and M_4, implying that $1 \leq y_1 \leq M_1,\ 1 \leq y_2 \leq M_2$, and so on. In total, there are $M_1 \times M_2 \times M_3 \times M_4$ possible combinations of responses (or response patterns). We denote the probability of occurrence of a specific combination of scores on these 4 variables by $P(y_1, y_2, y_3, y_4)$. LC analysis involves defining a specific type of model for these probabilities for response patterns.

The key assumption of any type of LC model is that each observation is a member of one and only one of *C* latent (unobservable) classes (i.e., Goodman, 1974a). Worded differently, the population of interest contains *C* subgroups, but we do not know to which subgroup an individual belongs. More technically, the population is a *mixture* of C latent classes. Let us look at the statistical implication of this mixture assumption for the case of a traditional LC model for four categorical responses. We will denote the latent variable by X and a particular latent class by c. The first basic equation of a LC model can now be defined as follows:

$$P(y_1, y_2, y_3, y_4) = \sum_{c=1}^{C} P(X = c) P(y_1, y_2, y_3, y_4 \mid X = c),$$

where $P(X = c)$ is the probability of belonging to latent class c and $P(y_1, y_2, y_3, y_4 \mid X = c)$ is the probability of having the set of responses concerned given that one belongs to latent class c. The above formula shows that each latent class or subgroup has its own (joint response) probability $P(y_1, y_2, y_3, y_4 \mid X = c)$ and that the overall probability $P(y_1, y_2, y_3, y_4)$ for the total population is obtained as a weighted average of the former using the class proportions $P(X = c)$ as weights. As an example, suppose we have 2 latent classes (C =2) and the probability of belonging to class 1 and 2 equals .6 and .4, respectively. Moreover, the probability of say response pattern (1,2,1,1) equals .2 in the first class and .05 in the second class. The overall probability of this response pattern will then equal $.6 \times .2 + .4 \times .05 = .14$.

The second key assumption of a traditional LC model is that *local independence* exists between the observed (manifest) variables. That is, conditional on latent class membership, the observed variables are mutually independent of each other. Using the same example with 4 variables, this can be expressed in an equation as follows:

$$P(y_1, y_2, y_3, y_4 \mid X = c) = P(y_1 \mid X = c) P(y_2 \mid X = c) P(y_3 \mid X = c) P(y_4 \mid X = c), \qquad (1)$$

where $P(y_1 \mid X = c)$ denotes the probability of giving response y_1 on the first observed variable given that one belongs to latent class c, $P(y_2 \mid X = c)$ is the corresponding conditional response probability for the second observed variable, and so on. As can be seen, conditional on the class one belongs to, the probability for a specific combination of 4 responses is a product of the probabilities of each of the separate responses, which is, in fact, the definition of (conditional) independence. This can be illustrated with a numerical example. Suppose that for latent class 1 the probability of choosing the first, second, first, and first response category on variables 1 to 4 equal

.8, .4, .9, and .7, respectively, then the probability for the full response pattern, $P(1,2,1,1 | X = 1)$, will be equal to $.8 \times .4 \times .9 \times .7 = .2$.

Combining the above mixture and local independence equations into a single equation yields the full formula for a traditional LC model for 4 observed variables; that is,

$$P(y_1, y_2, y_3, y_4) = \sum_{c=1}^{C} P(X = c) P(y_1 | X = c) P(y_2 | X = c) P(y_3 | X = c) P(y_4 | X = c) .$$

A more general formulation for any number of variables can be obtained as follows:

$$P(y_1, ..., y_J) = \sum_{c=1}^{C} P(X = c) \prod_{j=1}^{J} P(y_j | X = c) .$$

Here, we denote the number of observed variables by *J* and use the index *j* to refer to a particular observed variable. Note that $\prod_{j=1}^{J} P(y_j | X = c)$ is the shorthand notation for the product of the class-specific response probabilities for all *J* variables.

A LC model can be depicted graphically in terms of a path diagram (or a graphical model) in which manifest variables are not connected to each other directly, but indirectly through the common source *X*. The latent variable is assumed to explain all of the associations among the manifest variables. A goal of traditional LC analysis is to determine the smallest number of latent classes *C* that is sufficient to explain away (account for) the associations (relationships) observed among the manifest variables.

The analysis typically begins by fitting the C=1 class baseline model (H_0), which specifies mutual independence among the variables. Model H_0:

$$P(y_1, y_2, y_3, y_4) = P(y_1) P(y_2) P(y_3) P(y_4) .$$

Assuming that this *null* model does not provide an adequate fit to the data, a LC model with C=2 classes is then fitted to the data. This process continues by fitting successive LC models to the data, each time adding another dimension by incrementing the number of classes by 1, until the simplest model is found that provides an adequate fit.

2.1 Assessing Model Fit

Several complimentary approaches are available for assessing the fit of traditional LC models and determining the required number of latent classes. A widely-used approach utilizes the likelihood-ratio chi-squared *goodness-of-fit statistic* L^2 to assess the extent to which estimates for the expected cell frequencies according to the specified LC model, $\hat{\mu}_{y_1,\ldots,y_j}$, differ from the corresponding observed frequencies, $n_{y_1,\ldots,y_J}$:

$$L^2 = 2\sum n_{y_1,\ldots,y_J} \ln(n_{y_1,\ldots,y_J} / \hat{\mu}_{y_1,\ldots,y_J}),$$

where the sum is over all cells in the analyzed frequency table. The estimated frequencies $\hat{\mu}_{y_1,\ldots,y_j}$ are obtained by multiplying the estimated values of the probabilities $P(y_1,\ldots,y_J)$ for the response patterns by the sample size *N*. As an alternative to L^2, we may also use the Pearson or the Cressie-Read chi-squared statistic.

A model fits the data if the value of L^2 is sufficiently low to be attributable to chance (within normal statistical error limits --generally, the .05 level). In the case that $\hat{\mu} = n$ for each cell, the model fit will be perfect and L^2 equals zero. To the extent that the value for L^2 *exceeds* 0, the L^2 measures lack of model fit, quantifying the amount of association (non-independence) that remains unexplained by that model. When *N* is sufficiently large, L^2 follows a chi-square distribution, and as a general rule[1], the number of degrees of freedom (*df*) equals the number of cells in the full multi-way table minus 1 minus the number of distinct parameters *Npar*. For example, in the case of 4 categorical variables, the number of cells equals $M_1 \times M_2 \times M_3 \times M_4$ and the number of free parameters is:

$$Npar = (C-1) + C[(M_1-1) + (M_2-1) + (M_3-1) + (M_4-1)].$$

Note that *Npar* is obtained by counting the *C*-1 distinct LC probabilities, and for each latent class, the M_1-1 distinct conditional response probabilities associated with the categories of variable y_1, the M_2-1 distinct conditional probabilities associated with y_2, etc. Since probabilities sum to 1,

[1] According to the general rule, if it turns out that $df < 0$, the model is not identifiable, which means that unique estimates are not available for all parameters. For example, for variables with 2 categories, $df = -4$ for $T = 4$, which means that the 4-class model is not identifiable. In some cases however, this general counting rule may yield $df > 0$, yet the model may still not be identifiable. For example, Goodman (1974a) shows that in this situation of 4 dichotomous variables, the 3-class model is also unidentifiable despite the fact that the counting rule yields $df = 1$. See also note 3.

the probability associated with one category of each variable is redundant (and hence not counted as a *distinct* parameter): it can be obtained as one minus the sum of the others.

In situations involving a sparse frequency table (when expected cell frequencies are small), the theoretical chi-squared distribution should not be used to compute the p-value because the sampling distribution of L^2 would not be well approximated. Instead, the bootstrap approach can be used to estimate p (Langeheine, Pannekoek, and Van de Pol, 1996). Sparseness occurs when a LC model is specified for more than a few observed variables or when the sample size is very small. In both cases, the total number of cells in the resulting multi-way frequency table will be large relative to the sample size, resulting in many cells with low expected frequencies. This situation is illustrated below with a data example where the sample size is rather small.

An alternative approach to assessing model fit in the case of a sparse table, but also in other situations, utilizes an *information criterion* weighting both model fit and parsimony. Such measures, like AIC and BIC, are especially useful in comparing models. The most widely used in LC analysis is the BIC statistic which can be defined as: $BIC_{L^2} = L^2 - \ln(N)\, df$ (Raftery, 1986). A model with a lower BIC value is preferred over a model with a higher BIC value. A more general definition of BIC is based on the log-likelihood (LL) and the number of parameters (*Npar*) instead of L^2 and *df*; that is,

$$BIC_{LL} = -2\ LL + \ln(N)\ Npar$$

Again, a model with a lower BIC value is preferred over a model with a higher BIC value.[2]

If the baseline model (H_0) provides an adequate fit to the data, no LC analysis is needed, since there is no association among the variables to be explained. In most cases, however, H_0 will not fit the data in which case $L^2(H_0)$ can serve as a baseline measure of the total amount of association in the data. This suggests a 3rd approach for assessing the fit of LC models by comparing the L^2 associated with LC models for which $C>1$ with the baseline value $L^2(H_0)$ to determine the *percent reduction in* L^2. Since the total association in the data may be quantified by $L^2(H_0)$, the percent reduction measure represents the total association explained by the model. This less formal approach can complement the more statistically precise L^2 and BIC approaches.

As an example of how these measures are used, suppose that the L^2 suggests that a 3-class model falls short of providing an adequate fit to some data (say p = .04) but explains 90% of

[2] The two formulations of BIC differ only with respect to a constant. More precisely, BIC_{L^2} equals BIC_{LL} minus the BIC_{LL} corresponding to the saturated model. Note that LL equals the sum of $n_{y_1,\ldots,y_j}$ $\ln(P(y_1,\ldots, y_J))$ across the observed data patterns.

the total association. Moreover, suppose a 4-class is the simplest model that fits according to the L^2 statistic but that this model only explains 91% of the association. In this case, it may be that on practical grounds the 3-class model is preferable since it explains almost as much of the total association.

Below, we discuss two additional types of statistics. The first concerns *bivariate residuals*, which are Pearson chi-squared goodness-of-fit statistics for two-way marginal tables indicating how well a LC model picks up the association between pairs of variables (rather than the associations between all variables). The second type concerns *classification statistics*, which are not fit measures but which instead quantify how certain we are about the individuals' class memberships if we wish to use the LC model as a tool for clustering.

<u>Example: survey respondent types</u>

We will now consider a first example that illustrates how these tools are used in practice. It is based on the analysis of 4 variables from the 1982 General Social Survey given by McCutcheon (1987) to illustrate how traditional LC modeling can be used to study the different types of survey respondents. Two of the variables ascertain the respondent's opinion regarding (y_1) the purpose of surveys and (y_2) how accurate they are, and the others are evaluations made by the interviewer of (y_3) the respondent's levels of understanding of the survey questions and (y_4) cooperation shown in answering the questions. McCutcheon initially assumed the existence of 2 latent classes corresponding to 'ideal' and 'less than ideal' types.

The study included separate samples of white and black respondents. Beginning with an analysis of the white sample, McCutcheon later included data from the black sample to illustrate a 2-group LC analysis. We will use these data to introduce the basics of traditional LC modeling and to illustrate several recent developments that have been made over the past decade. These include allowing for specific local dependencies (section 3.1), the usage of LC factor models (section 3.2), and the inclusion of covariates as well as the methodology for making multi-group comparisons (sections 3.3 and 3.4).

[INSERT TABLE 1 ABOUT HERE]

Traditional exploratory LC analysis begins by fitting the null model H_0 to the sample of white respondents. Since $L^2(H_0) = 257.3$ with *df* = 29 (see Table 1), the amount of association (non-independence) that exists in these data is too large to be explained by chance, so the null model must be rejected ($p < .001$) in favor of C>1 classes.

Next, we consider McCutcheon's 2-class model (H_1). For this model, the L^2 is reduced to 79.5[3], a 69.1% reduction from the baseline model, but still much too large to be acceptable with $df = 22$. Thus, we increment C by 1 and estimate model H_{2C}, the 3-class model. This model provides a further substantial reduction in L^2 to 22.1 (a 91.5% reduction over the baseline) and also provides an adequate overall fit ($p>.05$). Table 1 shows that the 4-class LC model provides some further improvement. However, the BIC statistic, which takes parsimony into account, suggests that the 3-class model is preferred over the 4-class model (see Table 1).

[INSERT TABLE 2 ABOUT HERE]

The parameter estimates obtained from the 3-class model are given in the left-most portion of Table 2. The classes are ordered from largest to smallest. Overall, 62% are estimated to be in class 1, 20% in class 2 and the remaining 18% in class 3. Analogous to factor analysis where names are assigned to the factors based upon an examination of the 'factor loadings', names may be assigned to the latent classes based upon the estimated conditional probabilities. Like factor loadings, the conditional probabilities provide the measurement *structure* that defines the latent classes.

McCutcheon assigned the name 'Ideal' to latent class 1, reasoning as follows:
"The first class corresponds most closely to our anticipated ideal respondents. Nearly 9 of 10 in this class believed that surveys 'usually serve a good purpose'; 3 of 5 expressed a belief that surveys are either 'almost always right' or 'right most of the time'; 19 of 20 were evaluated by the interviewer as 'friendly and interested' during the interview; and nearly all were evaluated by the interviewer as having a good understanding of the survey questions." He named the other classes 'Believers' and 'Skeptics' based on the interpretations of the corresponding conditional probabilities for those classes.

2.2 Testing the Significance of Effects

The next step in a traditional LC analysis is to delete from the model any variable that does not exhibit a significant difference between the classes. For example, to test whether to delete variable

[3] This value differs slightly from the value 79.3 reported in McCutcheon (1987) because our models include a Bayes constant set equal to 1 in order to prevent boundary solutions (estimated model probabilities equal to zero). For further information on Bayes constants see the technical appendix of the Latent GOLD manual (Vermunt and Magidson 2000 or www.latentclass.com).

y_1 from a *C*-class model, one would test the null hypothesis that the distribution over the M_1 categories of y_1 is identical within each class *c*:

$$P(y_1 \mid X = 1) = P(y_1 \mid X = 2) = ... = P(y_1 \mid X = C) \text{ for } y_1 = 1, 2, \ldots, M_1 .$$

In order to implement this test we parameterize the conditional response probabilities in terms of logit parameters (see, e.g., Haberman 1979; Formann, 1992; or Heinen, 1996):

$$P(y_1 \mid X = c) = \frac{\exp(\alpha^1_{y_1} + \beta^1_{y_1 c})}{\sum_{m=1}^{M_1} \exp(\alpha^1_m + \beta^1_{mc})} .$$

This logistic specification for class-specific response probabilities for variable y_1 -- with intercept parameters $\alpha^1_{y_1}$ and slope parameters $\beta^1_{y_1 c}$ -- can then be used to test the null hypothesis. Re-expressed in terms of the linear logistic associated with the y_1-*X* relationship, this yields:

$$\beta^1_{y_1 1} = \beta^1_{y_1 2} = ... = \beta^1_{y_1 C} = 0 \text{ for } y_1 = 1, 2, \ldots, M_1 ,$$

that is, the test of the assumption that the slope parameters are equal to 0.

One way to test for significance of the 4 indicator-class associations in our 3-class model is by means of a L^2 difference test, where ΔL^2 is computed as the difference between the L^2 statistics obtained under the *restricted* and *unrestricted* 3-class models respectively. The ΔL^2 values obtained by setting the association parameters corresponding to one of the indicators to zero were 145.3, 125.4, 61.3, and 101.1 for y_1, y_2, y_3, and y_4, respectively. These numbers are higher than of the corresponding Wald statistics, which took on the values 29.6, 8.4, 7.4, and 19.0. This is because the latter test is uniformly less powerful than the ΔL^2 statistic. Under the assumption that the unrestricted model is true, both statistics are distributed asymptotically as chi-square with $df = (M_j-1)*(C-1)$, where M_j denotes the number of categories in the nominal response variable concerned. The encountered values show that each of the 4 indicators included in the model is significantly related to class membership.

2.3 Classification

Typically, an additional step in a traditional LC analysis is to use the results of the selected model to classify cases into the appropriate latent classes. For any given observed response pattern ($y_1,...,y_J$), the posterior probability of belonging to class *c*, $P(X=c\,|\,y_1,...,y_J)$, can be obtained using Bayes theorem as follows:

$$P(X=c\,|\,y_1,...,y_J)=\frac{P(X=c)P(y_1,...,y_J\,|\,X=c)}{\sum_{c'=1}^{C}P(X=c')P(y_1,...,y_J\,|\,X=c')}=\frac{P(X=c)\prod_{j=1}^{J}P(y_j\,|\,X=c)}{\sum_{c'=1}^{C}P(X=c')\prod_{j=1}^{J}P(y_j\,|\,X=c')}$$

for *c*=1,2,…,*C*. Note that the denominator in this formula is simply $P(y_1,...,y_J)$.

Magidson and Vermunt (2001) and Vermunt and Magidson (2002) refer to this kind of approach as a LC *clustering* because the goal of classification into *C* homogeneous groups is identical to that of cluster analysis. In contrast to an ad hoc measure of distance used in cluster analysis to define homogeneity, LC analysis defines homogeneity in terms of probabilities. As indicated by Equation (1), cases in the same latent class are similar to each other because their responses are generated by the same probability distribution.

Typically, cases are assigned to the class for which the posterior probability is highest (i.e., the modal class). For example, according to the 3-class LC model, someone with response pattern y_1=1 (PURPOSE = 'good'), y_2=1 (ACCURACY = 'mostly true'), y_3=1 (UNDERSTANDING = 'good'), and y_4=1 (COOPERATION = 'interested') has posterior membership probabilities equal to 0.92, 0.08, and 0.00. This means that such a person is assigned to the first class.

The performance of a LC model as a tool for clustering – or more technically, the amount of class separation – can be quantified using classification statistics. The simplest statistics is the proportion of classification errors. Under modal class assignment, the proportion of classification errors corresponding to a response pattern ($y_1,...,y_J$) equals $1-\max P(X=c\,|\,y_j,...,y_J)$, a number equal to 0 when the posterior probability equals 1 for one class and 0 for the other classes, and equal to its maximum values of (*C*-1)/*C* when all classes are equally likely. The overall proportion of classification errors is computed by taking the average across response patterns; that is,

$$\text{Classification Errors}=\sum n_{y_1,...,y_J}[1-\max P(X=c\,|\,y_1,...,y_J)]/N\,.$$

Another type of classification statistics are R^2 like measures which indicate how well we can predict class membership using the responses individuals provide compared to the situation in which we simply ignore these responses. The most popular measure is the entropy-based R^2, which obtained as follows:

$$\text{Entropy } R^2 = \frac{\text{Entropy(null)} - \text{Entropy(model)}}{\text{Entropy(null)}}.$$

The null and model entropy for a specific response pattern equal $-\sum_{c=1}^{C} P(X=c) \ln P(X=c)$ and $-\sum_{c=1}^{C} P(X=c \mid y_1,...,y_J) \ln P(X=c \mid y_1,...,y_J)$, respectively. In the above formula, we fill in the averages of these across response patterns.

2.4 Graphical Displays

Since for any given response pattern (y_1, y_2, y_3, y_4) the C class membership probabilities sum to 1, only C-1 such probabilities are required as the probability of belonging to the remaining class can be obtained from the others. Hence, the class membership probabilities $P(X=c \mid y_1, y_2, y_3, y_4)$ can be used to position each response pattern in C-1 dimensional space, and for C=3, various 2-dimensional barycentric coordinate displays can be produced.

Rather than plotting every one of the many response patterns, instructive plots of the kind used in correspondence analysis can be produced, where points are plotted for each category of each variable as well as other meaningful aggregations of these posterior probabilities (Magidson and Vermunt, 2001).

[INSERT FIGURE 1 ABOUT HERE]

Figure 1 depicts the corresponding barycentric coordinate display under the 3-class LC model. Points are plotted for each category of each of the 4 variables in our example. Since these points contain information equivalent to the LC parameter estimates (Van der Heijden, Gilula and Van der Ark, 1999), this type of plot provides a graphical alternative to the traditional tabular display of parameter estimates and can yield new insights into data. Also displayed in Figure 1 are 2 additional aggregations associated with the response categories UNDERTANDING = 'good' and 'fair, poor' (y_3 =1,2) among those for whom COOPERATION = 'hostile/impatient' (y_4 =3).

The horizontal dimension of the plot corresponds to differences between McCutcheon's 'ideal' and 'believer' types (latent classes 1 and 2). We see that the categories of the variable y_3 tend to spread out along this dimension. Respondents showing 'good' understanding are most likely to belong to the ideal class (the corresponding symbol is plotted closest to the lower left vertex that represents class 1) while those showing only 'fair or poor' understanding are plotted closest to the lower right vertex which represents class 2.

Differences along the vertical dimension of the plot are best shown by the categories of y_1 and y_2. For example, respondents agreeing that the purpose of surveys is 'good' are plotted close to the lower left (class 1) vertex. Those who say 'it depends' are plotted somewhat midway between the class 1 and class 3 (top) vertex. Those who say 'it's a waste of time and $' are most likely to be in class 3 and are positioned near the top vertex. The fact that the positioning of categories for both y_1 and y_2 spread out over the vertical dimension suggests a high degree of association between these variables. In contrast, the categories of y_3 are spread over the horizontal dimension, suggesting that the association between y_3 and the 2 variables y_1 and y_2 is close to nil.

The categories of the variable y_4 form an interesting diagonal pattern. Respondents showing they are 'interested' in the questions are most likely to be in class 1 ('ideal') while those who are only 'cooperative' or exhibit 'impatience/hostility' are plotted closer to classes 2 and 3. This suggests the hypothesis that impatience and hostility may arise for either of 2 different reasons – 1) disagreement that surveys are accurate and serve a good purpose (indicated by the vertical dimension of the plot) and/or 2) lack of understanding (indicated by the horizontal dimension).

The additional points plotted deal with the relationship between variables y_3 and y_4. The positioning of these points suggest that among impatient/hostile respondents, those who show good understanding of the questions tends to be more in class 3 while those whose understanding is Fair/Poor tend to be about equally likely to be in classes 2 or 3.

We will revisit these data and obtain further insights later when we examine an alternative *nontraditional* 2-dimensional LC model, the 2-factor LC model.

Example: sparse multi-rater agreement data

We next consider an example with sparse frequency table where 7 pathologists each classified 118 slides as to the presence or absence of carcinoma in the uterine cervix (Landis and Koch, 1977), This data set was also analyzed by Agresti (2002). LC modeling will be used here to estimate the false positive and false negative rates for each pathologist and to use multiple ratings to distinguish between slides that indicate carcinoma and those that don't (for similar medical applications see Rindskopf and Rindskopf, 1986; Uebersax and Grove, 1990). The second column of Table 3 shows that the raters vary from classifying only about 1 of every 5 slides as positive (Rater *4*) to classifying more than 2 of every 3 as positive (Rater *2*). The next two columns indicate for which percentage of slides the ratings agree among 5 or more and 6 or more raters. This information shows that agreement is highest among raters *3*, *1*, *7* and *5*.

[INSERT TABLE 3 ABOUT HERE]

As a starting point Agresti (2002) formulated a model containing 2 latent classes, in an attempt to confirm the hypothesis that slides are either 'true positive' or 'true negative'. The assumption of local independence in the 2-class model means that rater agreement is caused solely by the differing characteristics between these 2 types of slides. That is, given that a slide is in the class of 'true positive' ('true negative') any similarities and differences between raters represent pure error. However, in his analysis of these data he found that 3-classes were necessary to obtain an acceptable fit.

While there are $2^7 = 128$ possible response patterns, because of the large amount of inter-rater agreement, 108 of these patterns were not observed at all. As mentioned above, sparse data such as this causes a problem in testing model fit because the L^2 statistic does not follow a chi-square distribution. For this reason, Agresti simply alluded to the obvious discrepancy between the expected frequencies estimated under the 2-class model and the observed frequencies and speculated that this model does not provide an adequate fit to these data. He then compared estimates obtained from the 3-class model, and suggested that the fit of this model was adequate.

[INSERT TABLE 4 ABOUT HERE]

We report the bootstrap p-value in Table 4, which confirms Agresti's speculation that the fit of the 2-class model *is* poor and that of the 3-class model is adequate. It also shows that the 3-class model is preferred over the 4-class model according to the BIC criteria.

The parameter estimates obtained with the 3-class model are given in the middle portion of Table 3. The largest class (44%) refers to slides that all pathologists (except for *4* and *6*) almost always agree show carcinoma ('true positive'). Class 2 (37%) refers to slides that all pathologists (except occasionally *2*) agree shows no carcinoma. The remaining class of slides (18%) shows considerable disagreement between pathologists – *2*, *5* and *7* usually diagnose carcinoma while *3*, *4* and *6* rarely do and *1* diagnoses carcinoma half the time.

If we assume that class 1 represents cases of true carcinoma, the results reported in Table 3 show that those pathologists who rated the *fewest* slides as positive (*4* and *5*), have the highest false *negative* rates (42% and 53% respectively, highlighted in bold). Similarly, under the assumption that class 2 represents cases free from carcinoma, the results show that the pathologist who rated the *most* slides as positive, pathologist *2*, shows a false positive rate (15%) that is substantially larger than the other pathologists.

The traditional model fitting strategy requires us to reject our 2-class hypothesis in favor of a 3-class alternative where the 3rd latent class consists of slides that can not be classified as either 'true positive' or 'true negative' for cancer. Next we consider some nontraditional LC models which provide classification of each slide according to its likelihood of carcinoma. In particular, we will show that a 2-factor LC model provides an attractive alternative where factor 1 classifies all slides as either 'true positive' or 'true negative', and factor 2 classifies slides according to a tendency for ratings to be biased towards false positive or false negative error.

3. NONTRADITIONAL LATENT CLASS MODELING

Rejection of a traditional *C*-class LC model for lack of fit means that the local independence assumption does not hold with *C* classes. In such cases the traditional LC model fitting strategy is to fit a *C*+1 class model to the data. In both of our examples, theory supported a 2-class model but since this model failed to provide an adequate fit we formulated a 3-class model. In this section we consider some alternative strategies for modifying a model. In both cases we will see the nontraditional alternatives lead to models that are more parsimonious than traditional models as well as models that are more congruent with our initial hypotheses. The alternatives considered are:

1. adding 1 or more direct effects

2. deleting 1 or more items
3. increasing the number of latent variables

Alternative #1 is to include 'direct effect' parameters in the model (Hagenaars, 1988) that account for the residual association between the observed variables that is responsible for the local dependence. This approach is particularly useful when some external factor, unrelated to the latent variable, creates irrelevant association between two variables. Examples of such external factors include similar question wording used in two survey items, as well as two raters using the same incorrect criterion in evaluating slides.

Alternative #2 also deals with the situation where 2 variables are responsible for some local dependency. In such cases, rather than add a direct effect between two variables, it may make more sense to eliminate the dependency by simply deleting one of the two items. This variable reduction strategy is especially useful in situations where there are *many* redundant variables.

Alternative #3 is especially useful when a group of several variables account for a local dependency. Magidson and Vermunt (2001) show that by increasing the dimensionality through the addition of latent variables rather than latent classes, the resulting LC *factor* model often fits data substantially better than the traditional LC cluster models having the same number of parameters. In addition, LC factor models are identified in some situations where the traditional LC model is not[4].

In the next section, we introduce a diagnostic statistic called the *bivariate residual* (BVR) and illustrate its use to develop some nontraditional alternative models for our two data examples. The BVR helps pinpoint those *bivariate* relationships[5] that fail to be adequately explained by the LC model and can help determine which of the 3 alternative strategies to employ. We will see that even in situations where the L^2 statistic reports that the model provides an adequate *overall* fit, the fit in one or more 2-way tables may not be adequate and may indicate a flaw or weakness in the model.

3.1 Bivariate Residuals and Direct Effects

[4] For example, with 4 dichotomous variables, a LC 2-factor model (comprised of 4 latent classes) is identified whereas a traditional 3-class model is not (Goodman, 1974a).

[5] Traditional factor analysis, through the assumption of multivariate normality, limits its focus to bivariate relationships (i.e., the correlations), since higher-order relationships are assumed not to exist. In contrast, LC models do not make strict distributional assumptions, and hence attempt to explain higher-order associations as well. Nevertheless, the 2-way (bivariate) associations are generally the most prominent, and the ability to pinpoint specific 2-way tables where lack of fit may be concentrated can be useful in suggesting alternative models.

A formal measure of the extent to which the observed association between 2 variables is reproduced by a model is given by the BVR statistic (Vermunt and Magidson, 2001). Each BVR corresponds to a Pearson X^2 statistic (divided by the degrees of freedom) where the observed frequencies in a 2-way cross-tabulation of the variables are contrasted with those expected counts estimated under the corresponding LC model[6]. A BVR value substantially larger than 1 suggests that the model falls somewhat short of explaining the association in the corresponding 2-way table.

[INSERT TABLE 5 ABOUT HERE]

Example: survey respondent types (continued)

Table 5 reports BVRs for each variable pair under each of several models estimated in our first example. Since model H_0 corresponds to the model of mutual independence, each BVR for this model provides a measure of the overall association in the corresponding observed 2-way table; that is, each BVR equals the usual Pearson X^2 statistic used to test for independence in the corresponding 2-way table divided by the degrees of freedom. The results show that except for the non-significant relationships in the $\{y_1,y_3\}$ and $\{y_2,y_3\}$ tables, all of the remaining BVRs are quite large, attesting to several significant associations (local dependencies) that exist among these variables. The BVR is especially large for $\{y_1,y_3\}$ and for $\{y_3,y_4\}$. For example, in Table $\{y_3,y_4\}$, a Pearson chi-square test confirms that the observed relationship is highly significant ($X^2 = 86.8$, $df=2$, $p<.001$; BVR = 86.8/2 = 43.4).

Under the 2-class model (H_1), note that the BVRs are all near or less than 4 except for one very large value of 32.3 for $\{y_3,y_4\}$. This suggests that the overall lack of fit for this model can be traced to this single large BVR. The traditional way to account for the lack of fit is by adding another latent class. Table 5 shows that after the addition of a 3rd class, the BVR for $\{y_3,y_4\}$ under the 3-class model H_{2C} is at an acceptable level (BVR = 2.4).

Below, we consider the alternative approach of adding a 'direct effect' to the model to account for the residual correlation. In addition, we consider use of the 2-factor LC model and further explore the differences between the 3 and 4-class models.

[INSERT TABLE 6 ABOUT HERE]

[6] These residuals are similar to Lagrange-multiplier or Score test statistics. A difference is that they are limited information fit measures: dependencies with parameters corresponding to other items are not taken into account.

Example: sparse multi-rater agreement data (continued)

Turning now to our second example, Table 6 shows that all of the BVRs under the 1-class model of mutual independence (model H_0) are very large[7], indicating that the amount of agreement between each pair of raters is highly significant. Under the 2-class model many BVRs remain large. While the 3-class model provides an acceptable *overall* fit to these data, again we see that there is a single BVR that remains somewhat large – BVR = 4.5 for raters *4* and *5*, the 2 pathologists who rated the fewest slides positive (recall Table 3). This larger BVR suggests that raters *4* and *5* may be using some rating criterion not shared by the other raters.

To account for this larger residual association, we will use nontraditional alternative #1 and modify the 3-class model by adding the $\{y_4,y_5\}$ direct effect parameter(s) $\alpha^{45}_{y_4 y_5}$ into the model (Hagenaars, 1988; for a slightly different formulation, see Uebersax, 1999). Formally, this new model H_{2C+} is expressed as:

$$P(y_1,...,y_7) = \sum_{c=1}^{C} P(X=c)P(y_1 \mid X=c)P(y_2 \mid X=c)P(y_3 \mid X=c)P(y_4,y_5 \mid X=c)P(y_6 \mid X=c)P(y_7 \mid X=c)$$

where the probabilities $P(y_4,y_5 \mid X=c)$ are constrained as follows:

$$P(y_4,y_5 \mid X=c) = \frac{\exp(\alpha^4_{y_4} + \alpha^5_{y_5} + \alpha^{45}_{y_4 y_5} + \beta^4_{y_4 c} + \beta^5_{y_5 c})}{\sum_{m_4=1}^{M_4}\sum_{m_5=1}^{M_5}\exp(\alpha^4_{m_4} + \alpha^5_{m_5} + \alpha^{45}_{m_4 m_5} + \beta^4_{m_4 c} + \beta^5_{m_5 c})}$$

By relaxing the local independence assumption between raters *4* and *5*, model H_{2C+} is able to account for excessive association between *4* and *5* that is not explainable by the latent classes. The ΔL^2 test shows that inclusion of the direct effect parameter provides a significant improvement over the traditional model H_{2C} (ΔL^2 = 17.7 – 11.3 = 6.4; p = .01).

From a practical perspective, models H_{2C} and H_{2C+} do not differ much as both models assign the 126 slides to the same classes under the modal assignment rule. This occurs despite the fact that model H_{2C+} gives *4* and *5* less weight than model H_{2C} during the computation of the posterior probabilities. The primary benefit of model H_{2C+} is to suggest the possibility that raters *4* and *5* share a bias when evaluating class 1 slides, those slides that *4* and *5* often rate negative but that the other pathologists almost always rate positive (recall Table 3). The implication of

[7] The smallest BVR under model H_0 is 20.8 which occurs in table $\{y_5,y_6\}$.

including the direct effect is that model H_{2C+} provides higher predictions of *agreement* between *4* and *5* than model H_{2C} on class 1 slides[8].

Returning to the first data example for a moment, we might now expect to find similar insights by the inclusion of the direct effect parameters, $\alpha^{34}_{y_3 y_4}$, in the 2-class model. Table 1 shows that this model (H_{1C+}) provides a good fit to the data. However, under this model, the parameter measuring the contribution of y_3 to the latent classes is no longer significant and therefore y_3 can be deleted from the LC model completely. As this amounts to deleting an association simply because it could not be explained by a model with 2 latent classes, alternative #1 does not provide a desirable solution here.

3.2 LC Factor Models

Next we consider alternative approach #3 where we utilize LC factor models to include more than one latent variable in the model. LC factor models were proposed as a general alternative to the traditional exploratory LC modeling by Magidson and Vermunt (2001). For both examples, the results (given in Table 1 and Table 4) show that a 2-factor model is preferable to the other models. We shall see that the 2-factor model is actually a restricted 4-class model. In both cases the fit is almost as good as the (unrestricted) 4-class solution, but is more parsimonious and parameterized in a manner that allows easier interpretation of the results.

LC factor models were initially proposed by Goodman (Goodman, 1974b) in the context of confirmatory latent class analysis. Certain traditional LC models containing 4 or more classes can be interpreted in terms of 2 or more component latent variables by treating those components as a joint variable (see e.g., McCutcheon 1987; Hagenaars 1990). For example, a latent variable *X* consisting of *C*=4 classes can be re-expressed in terms of 2 dichotomous latent variables F_1 = {1,2} and F_2 = {1, 2} using the following correspondence:

	F_2=1	F_2=2
F_1=1	X =1	X = 2
F_1=2	X =3	X = 4

[8] Since model H_{2C} assumes local independence, the expected probability of both raters agreeing that a given class 1 slide is free from cancer can be computed by multiplying the corresponding conditional probabilities. Using the estimates from Table 3, the probability of both agreeing that a class 1 slide is negative is .42 x .53 = .22, and similarly, the probability of both agreeing that it was positive is .59 x .47 = .28. In contrast, model H_{2C+} predicts higher probabilities (.31 and .35 respectively) for rater 4 and 5 *agreeing* in both cases. Under the assumption that class 1 slides are 'true positive', the results from model H_{2C+} mean that raters 4 and 5 both tend to share a bias towards committing a false negative error.

Thus, X=1 corresponds with F_1=1 and F_2=1, X=2 with F_1=1 and F_2=2, X=3 with F_1=2 and F_2=1, and F_2=4 with F_1=2 and F_2=2.

Formally, for 4 nominal variables, the 4-class LC model can be re-parameterized as a LC factor model with two dichotomous latent variables as follows:

$$P(y_1, y_2, y_3, y_4) = \sum_{c_1=1}^{2}\sum_{c_2=1}^{2} P(F_1 = c_1, F_2 = c_2)\prod_{j=1}^{4} P(y_j \mid F_1 = c_1, F_2 = c_2).$$

Magidson and Vermunt (2001) consider various restricted factor models. They use the term *basic* LC factor models to refer to certain LC models that contain 2 or more dichotomous latent variables that are mutually independent of each other and that exclude higher-order interactions from the conditional response probabilities. Such a model is analogous to the approach of traditional factor analysis where multiple latent variables are used to model multidimensional relationships among manifest variables.

It turns out that by formulating the model in terms of R mutually independent, dichotomous latent factors, the basic LC factor model has the same number of distinct parameters as a traditional LC model with R+1 classes. That is, the LC factor parameterization allows specification of a 2^R-class model with the same number of parameters as a traditional LC model with only R+1 classes! This offers a great advantage in parsimony over the traditional T- class model as the number of parameters is greatly reduced by natural restrictions.

As mentioned previously, the basic 2-factor model provides an excellent fit to both of our example data sets. For the first example, Table 1 shows that this model (model H_{2F}) is preferred over any of the LC cluster models according to the BIC. In addition, this model explains all bivariate relationships in the data (see Table 5). We will interpret the results from this model in the next section in conjunction with a more extensive analysis including both the white and black sample.

<u>Example: sparse multi-rater agreement data (continued)</u>

Regarding our second example, Table 4 shows that the basic 2-factor model is preferred over all the other models according to the BIC criteria. The right-most portion of Table 3 provides the parameter estimates[9] that we used to name the factors. These are joint latent class and conditional

[9] The 2-factor model in Tables 3 was further restricted by setting the effect of indicator C on factor 2 to zero since this effect was not significant.

response probabilities for combinations of factor levels. We assigned the name 'True -' and 'True +' to levels 1 and 2 of factor 1 respectively. Each of these levels is split again into 2 levels by factor 2, which we named 'tendency towards ratings bias'. We named the 2 levels of factor 2 'tend to – bias' and 'tend to + bias' respectively.

Comparing the four factor cells (right-most portion of Table 3) to the classes in the 3-class model (middle portion of Table 3) we see the following similarities. First, note that class 1 of the 3-class solution (representing 44% of the slides mostly rated +) corresponds primarily to factor 1 level 2 slides (those named 'True +'), which account for 46% of all slides. These 'True +' slides are divided according to factor 2 into cell (2,1) accounting for 30% of all slides and cell (2,2) accounting for 16% of the slides. Note that the former slides show a clear tendency towards a false negative error, especially among raters 4 and 5.

Next, notice the similarity between class 2 of the 3-class solution, representing 37% of the slides rated mostly negative, and factor cell (1,1) accounting for 36% of the slides rated mostly negative. In addition, from Table 3 we can also see the strong similarity between class 3 of the 3-class solution and factor cell (1,2), identified in the table as 'True –' slides that are prone to 'false +' error, especially by raters 1, 2, 5 and 7.

In conclusion, we have shown that the 2-factor LC model fits better than the traditional 3-class model and offers two substantive advantages. First, it provides a clear way to classify slides as 'True +' of 'True –'. Second, it provides a further grouping of slides that may be useful in pinpointing the reasons for rater disagreement. Of course, whether factor 1 *actually* distinguishes between 'True –' and 'True +' and whether the error characterization given by factor 2 is accurate are important questions that could be addressed in future research.

3.3 Multi-group Models

Multi-group LC models can be used to compare models across groups. A completely unrestricted multi-group LC model, referred to by Clogg and Goodman (1984) as the model of complete heterogeneity, is equivalent to the estimation of a separate *C*-class LC model for each group. The fit of such a model can be obtained by simply summing the L^2 values (and corresponding degrees of freedom) for the corresponding models in each group.

Let *G* denote a categorical variable representing membership in group *g*. The model of *complete heterogeneity* is expressed as (model M_{2C}):

$$P(y_1, y_2, y_3, y_4 \mid G = g) = \sum_{c=1}^{C} P(X = c \mid G = g) \prod_{j=1}^{4} P(y_j \mid X = c, G = g).$$

Example: survey respondent types (continued)

The second part of Table 1 provides the results of repeating our example 1 analyses for the sample of *black* respondents. These results turn out to be very similar to those obtained for the white respondents (see first part Table 1). As in our analysis for the white sample we again reject the 1 and 2-class models in favor of 3 classes in order to obtain a model that provides an overall fit to the data that is adequate. The right-most portion of Table 2 presents the parameter estimates obtained from the 3-class model (model H'_{2C}) as applied to the sample of blacks. As in our earlier analysis the classes are ordered from largest to smallest.

In comparing results across these two groups, it is important to be able to interpret the 3 classes obtained from the black respondents as representing the same latent constructs ('ideal', 'believers' and 'skeptics') as in our analysis of the white respondents. Otherwise, any between-group comparisons would be like comparing 'apples' with 'oranges'. While it is tempting to interpret class 1 for both samples as representing the 'ideal' respondents, this is not appropriate without first restricting the measurement portion of the models (the conditional probabilities) to be equal. These restrictions are accomplished using the model of *partial* homogeneity (model M_{2CR}):

$$P(y_1, y_2, y_3, y_4 \mid G = g) = \sum_{c=1}^{C} P(X = x \mid G = g) \prod_{j=1}^{4} P(y_j \mid X = t) \qquad (3)$$

[INSERT TABLE 7 ABOUT HERE]

Estimates from this model are given in the left-most portion of Table 7. The third part of Table 1 compares the fit of the unrestricted model M_{2C} and restricted model M_{2CR}. The ΔL^2 statistic can be used to test the restrictions made under model M_{2CR}. Since $\Delta L^2 = 9.0$ with df=18 is *not* significant, we are free to use this restricted model for our group comparisons.

The model of complete homogeneity (model M_{2CRR}) imposes the further restriction that the latent class probabilities across the groups are identical: $P(X = c \mid G = 1) = P(X = c \mid G = 2)$, for c = 1,2,3. Since these restrictions yield a significant increase in L^2, we reject the model of complete homogeneity in favor of the model of partial homogeneity and conclude that there are significant differences in latent class membership between the white and black samples.

Table 1 also includes results obtained from the LC factor model counterparts to the models of complete heterogeneity and partial heterogeneity. Since these models contain 2 dichotomous and independent factors, they contain exactly the same number of parameters as the 3-class models M_{2C} and M_{2CR}. The lower-part of Table 1 shows that these models fit better than the corresponding LC cluster models according to the BIC criteria. Also the smaller BVRs than the LC cluster counterpart confirm that the LC factor model fits the data better.

The parameter estimates from the 2-factor model M_{2FR} are presented in the right-most portion of Table 7. These are marginal latent class and conditional response probabilities for factors F_1 and F_2, which are obtained by summing over the other factor. Note that variable y_4 is strongly related to both factors F_1 and F_2. That is, respondents at level 1 of each factor have a higher probability (.90 or .91) of being 'interested' than those at level 2. Variables y_1 and y_2 relate only to factor F_1 and variable y_3 relates only to factor F_2. That is, for factor F_1, those at level 1 are substantially more likely to agree that surveys serve a good purpose and are accurate than those at level 2, but the 2 levels are about equal in showing a good understanding of the questions. For factor F_2, level 1 shows good understanding while level 2 does not.

Moreover, Table 7 shows that group differences exist primarily with respect to factor 2 (observed group differences on factor F_1 are not significant). Black respondents are twice as likely as whites to be at level 2 of factor 2 (30% vs. 15%). These results allow us to formulate a more rigorous test of our earlier hypothesis that cooperation may be due to two separate factors – one associated with the belief that surveys serve a good purpose and are accurate (as assessed by LC factor 1), the second related to understanding the questions (as assessed by LC factor 2).

Before concluding this section, we note that thus far we have treated the trichotomous variables COOPERATE (y_1) and PURPOSE (y_3) as nominal. Alternatively, they can be treated as ordinal, which serves to simplify the model by reducing the number of parameters. The most straightforward approach is to restrict the logit parameters by using uniform scores for the categories of y_1 and y_3, implying the following constraints: $\beta^{11}_{y_1c_1} = \beta^{11}_{c_1} y_1$ and $\beta^{12}_{y_1c_2} = \beta^{12}_{c_2} y_1$ (see, e.g., Formann, 1992; or Heinen, 1996).

The use of these restrictions in our example increased the L^2 by very little, indicating that variables y_1 and y_3 may in fact be treated as ordinal. In the next section, we present the results of a modified 2-factor model where variables y_1 and y_3 are treated as ordinal.

3.4 Covariates

The parameters in the traditional LC model consist of unconditional and conditional probabilities. The conditional probabilities comprise the measurement portion of the model. They characterize the distribution among the observed variables (indicators) conditional on the latent classes. The *unconditional* probabilities describe the distribution of the latent variable(s). In order to obtain improved description/ prediction of the latent variable(s), a multinomial logit model is used to express these probabilities as a function of one or more exogenous variables ***Z*** called covariates (Dayton and McReady, 1988).

The multi-group model described in the previous section is an example of the use of a single nominal covariate (***Z*** = *G*). For example, the term $P(X = c \mid G = g)$ appearing Equation (3) can be expressed as:

$$P(X = c \mid G = g) = \frac{\exp(\gamma_c^0 + \gamma_{cg}^1)}{\sum_{c'=1}^{C} \exp(\gamma_{c'}^0 + \gamma_{c'g}^1)}.$$

While the latent variable(s) explain all of the associations among the indicators, associations between the covariates are *not* explained by the latent variables. This is what distinguishes the indicators from the covariates.

<u>Example: survey respondent types (continued)</u>

Regarding the interpretation of the 3-class solution, McCutcheon questioned whether some of the difference in latent class membership between black and white respondents might be explained by education, a question that falls outside the scope of traditional LC modeling. We address this question below by including *E*:EDUCATION as a second covariate in the 2-factor model -- ***Z*** = (*G*,*E*).

[INSERT TABLE 8 ABOUT HERE]

The model provides a good fit to the data. The results indicate that the effect of education *does* explain most, but not all, of the group (race) effect on factor F_2. The logit parameter estimates are given in Table 8, where nonsignificant estimates were set to zero. The multinomial model used for the covariate effects on F_2 was:

$$P(F_2 = c_2 \mid G = g, E = e) = \frac{\exp(\gamma_{c_2}^{02} + \gamma_{c_2 g}^{12} + \gamma_{c_2 e}^{22})}{\sum_{c=1}^{C_2} \exp(\gamma_c^{02} + \gamma_{cg}^{12} + \gamma_{ce}^{22})}.$$

The gamma parameters in Table 8 indicate that the higher the educational level the lower the score on factor F_2. The race effect is very weak: blacks have a slightly higher score on factor F_2 than whites.

[INSERT FIGURE 2 ABOUT HERE]

The results for this 2-factor restricted multi-group model are also displayed in the bi-plot display (Magidson and Vermunt, 2001) given in Figure 2. Like the barycentric coordinate display in Figure 1 we see that the horizontal axis, corresponding to factor F_2, is associated with UNDERSTANDING. Overall, respondents having a good understanding are highly likely to be at level 1 of factor F_2 while those with a Fair/Poor understanding are highly likely to be in level 2. The figure makes it clear that education is much more related to this factor than race. The vertical dimension is highly related to PURPOSE. Figure 2 shows more clearly than Figure 1 that COOPERATION is related to both factors. In particular, those rated as 'Impatient/hostile' tend to include 2 different types of respondents – those whose understanding is fair/poor as well as those who view the purpose of surveys as a 'waste of time and $'.

3.5 Three-step LC Analysis

Rather than including covariates directly within the estimated LC model, one may also use the following type of three-step approach:

1. Perform model selection and estimation in the usual way, thus without the inclusion of covariates;
2. Obtain class assignments W using the selected model from step 1, as explained in section 2.3;
3. Perform subsequent analyses with covariates or other types of external variables using the class assignments W of step 2.

Although this stepwise approach separating the LC analysis from the analyses one would like to do after the latent classes are constructed is very practical and intuitive, it is also problematic. More specifically, as a result of the classification errors introduced in step 2, it yields underestimated associations between external variables and latent classes. The larger the classification errors, the larger the bias in the estimates of these associations.

However, building upon the work by Bolck, Croon, and Hagenaars (2004), Vermunt (2010) proposed a solution to this problem. He showed how to perform a valid step-3 analysis by adjusting for the classification errors introduced in step 2. Basically, what happens is that a new

LC model is estimated in which the class assignments W are use as the single indicator with known conditional response probabilities $P(W = d \mid X = c)$. This adjusted step-3 analysis cannot only be used with covariates predicting class membership (via a logistic model), but also for investigating how classes differ with respect to a distal outcome variable (Bakk, Tekle, and Vermunt, 2013). Bakk and Vermunt (2016) recommended using the more robust BCH adjustment method for continuous distal outcomes (dependent variables), while for covariates and categorical distal outcomes the ML approach is preferred.

4. OTHER TYPES OF LATENT CLASS MODELS

Thus far, we have focused on the traditional LC modeling approach, including some important extensions such as covariates, several latent variables, and local dependencies. Some common characteristics of these models are that they serve as scaling methods or tools for dealing with measurement error, that indicators are nominal or ordinal, and that local independence between indicators is the primary model assumption. In this section, we discuss other types of LC models. They are not used as scaling tools, but as clustering methods, tools for dealing with unobserved heterogeneity, density estimation methods, or random-coefficients models (McLachlan and Peel, 2000). Moreover, indicators or dependent variables can be of scale types other than nominal or ordinal and local independence is no longer the basic model assumption. As we will see, in some cases there is only one indicator or dependent variable.

The next section presents simple mixture models for univariate distributions, with examples of mixtures of normals and mixtures of Poisson distributions. Then, we extend this basic model by including predictors, yielding what is called mixture regression or LC regression models. We present an example of a mixed linear regression model, and show how the method can deal with various types of repeated measurements. Special attention is given to the relationship with hierarchical or multilevel models. Then, we present another extension of the simple mixture model; that is, a mixture model for multivariate distributions. As will be shown, the resulting LC model can be seen as a model-based alternative to standard hierarchical clustering methods like K-means. We end with a short overview of LC methods that were not discussed in detail.

4.1 Simple Mixture Models

[INSERT FIGURE 3 ABOUT HERE]

Consider the histogram depicted in Figure 3. This generated data set of 1,000 cases is obtained from a population consisting of a mixture of two normal distributions. For 60 percent of the population, the variable of interest follows a normal distribution with a mean of 0 and a variance of 1, N(0,1); for the other 40 percent, the mean equals 3 and the variance 4, N(3,4). The normal curve that is drawn through the histogram shows that the resulting mixture is clearly not normally distributed.

A model that can be used to describe such a phenomenon is a finite mixture model (Everitt and Hand, 1981; McLachlan and Peel, 2000), which is a particular kind of LC model. The basic formula for a mixture of univariate distributions is

$$f(y \mid \vartheta) = \sum_{c=1}^{C} P(X = c) f(y, \phi_c) . \tag{4}$$

The left-hand side of Equation (4) indicates that we are interested in describing the distribution of a random variable *y*, which depends on a set of unknown parameters ϑ. The right-hand side contains two terms: $P(X = c)$ is the probability of belonging to latent class or mixture component *c* and $f(y \mid \phi_c)$ is the distribution of *y* within latent class *c* given some unknown parameters ϕ_c. The class-specific distribution of *y* is assumed to belong to a particular parametric family. Depending on the scale-type of *y*, this can, for instance, be a normal, Poisson, binomial, exponential, or gamma distribution. The summation on the right-hand side indicates that the distribution of *y* is a weighted mean of the class-specific distributions, where the latent class proportions serve as weights.

Mixture models like these have two important types of applications. The first is density estimation: complicated distributions can be approximated by a mixture of simple parametric distributions. Another important application type is clustering, in which case the class-specific parameters are used to define the clusters and the posterior membership probabilities are used to classify cases into the most appropriate cluster.

[INSERT TABLE 9 ABOUT HERE]

Table 9 presents test results for various models fitted to the data depicted in Figure 3. We estimated 1- to 4-class mixtures of normal distributions with equal and unequal within cluster-variances. As can be seen, the BIC measure identifies the correct model, the two-cluster model with unequal within-cluster variances, as best. The three-class model with equal within-cluster variances fits almost as well as, showing that a simpler parametric form can sometimes be compensated by a larger number of mixture components.

In the two-class model with unequal variances, the estimated probability of belonging to class one is .64. This class has an estimated mean of -0.03 and a variance of 1.01. The mean and variance of the other class equals 3.24 and 3.95. Note that these estimates are close to the population values we used to generate this data set.

[INSERT TABLE 10 ABOUT HERE]

Table 10 provides a data set taken from Dillon and Kumar (1994) that we will use as a second example. It gives the observed frequency distribution of the number of packs of hard-candy consumed by 456 respondents during the 7 days prior to the survey. Because the outcome variable is a count without a fixed maximum, it is most natural to assume that it follows a Poisson distribution. The table also reports the estimated frequency distribution obtained with a standard, or 1-class, Poisson model, as well as with a 3-class mixture Poisson model. As can be seen, the standard Poisson model does not fit the empirical distribution at all, while the 3-class Poisson describes the data almost perfectly. This shows that a mixture of simple parametric distributions can be used to describe a quite complicated empirical distribution.

Test results obtained when applying mixture Poisson models to the hard-candy data set show that models with 2 and 3 mixture components perform much better than the standard Poisson model. As is typical, there is a saturation point at which increasing the number of classes no longer increases the log-likelihood function: in this case, it occurs at 4 classes. The 3-cluster solution is the one that is preferred according to the BIC criterion.

The estimated latent class proportions in the 3-class model are 0.54, 0.28, and 0.18, and the Poisson rates are 3.48, 0.29, and 11.21. This means that we identified a small cluster of heavy users (more than 11 packs in 7 days), a cluster containing slightly more than a quarter of the respondents with almost no usage, and a large group of moderate users.

4.2 LC Regression Models

In the simple mixture models discussed above, it was assumed that the mean of the chosen parametric distribution differs across latent classes. This can also be expressed by specifying a linear regression model for the mean of the distribution of interest, μ_c, after applying some transformation or link function *g*(..) that depends on the scale type of the *y* variable. For the mean of a binomial or multinomial distribution, we use a logit transformation; for a Poisson mean, a log transformation; and for a normal mean, no transformation or an identity link. The regression model has the form

$$g(\mu_c) = \beta_{0c}.$$

As can be seen, this regression model contains only an intercept and this intercept is class-specific.

Let *w* denote a set of predictors or explanatory variables. Suppose we are no longer interested is the unconditional distribution of *y*, but in the conditional distribution of *y* given *w*, $f(y \mid \mathbf{w}, \phi_c)$. A natural way to express the dependency of *y* on *w* is by the inclusion of the set of predictors ***w*** in the right-hand side of the regression equation. In the case of a single predictor *w*, the resulting LC regression model (Wedel and DeSarbo, 1994) has the form

$$g(\mu_c) = \alpha_c + \beta_c w,$$

where α_c and β_c are the class-specific regression coefficients (intercept and slope).

[INSERT FIGURE 4 ABOUT HERE]

Figure 4 depicts a data set generated from a population consisting of two latent classes, with class-specific regression models equal to $\mu_1 = 1 + 3w$ and $\mu_2 = 0 + 1w$. It also compares the estimated *y* values for the two-class model (YLC2), with the standard, 1-class, regression model (YLC1). As can be seen, the description given by the standard regression model is very poor compared to the 2-class model. The LC regression modeling procedure has no problem identifying the two regression lines without pre-knowledge of class membership.

In a LC regression model, the latent variable is a predictor that interacts with the observed predictors, which means that it serves as a moderator variable. Compared to a standard regression model where all predictors are observed this basic LC regression model provides several useful functions. First, it can be used to weaken standard regression assumptions about the nature of the effects (linear, no interactions) and the error term (independent of predictors, particular distribution, homoskedastic). Second, it makes it possible to identify and correct for sources of unobserved heterogeneity. As explained below, this is especially useful in situations where there are repeated measurements or other types of dependent observations. Longitudinal data applications are sometimes referred to as LC or mixture growth models (each latent class has its own growth curve). Third, it can be used to detect outliers since these are cases for which the primary regression model does not hold.

An important application area for LC regression modeling is clustering or segmentation (Wedel and Kamakura, 1998). In particular, ratings- and choice-based conjoint studies are designed to identify subgroups (segments) that react differently to product characteristics, which is the same as saying that these groups have different regression coefficients. This type of application is illustrated in more detail below with an empirical example.

Example: repeated measurements or clustered observations

As explained below, the LC regression model can be viewed as a random-coefficients model that, similar to multilevel or hierarchical models, can take dependencies between observations into account. This extends the application of LC regression models to situations with repeated measurements or other types of dependent observations.

We will illustrate LC regression with repeated measurements using an application to longitudinal survey data. This is, therefore, an example of a LC growth model. The data set consist of 264 participants in the 1983 to 1986 yearly waves of the British Social Attitudes Survey (McGrath and Waterton, 1986). The dependent variable is the number of yes responses on seven yes/no questions as to whether it is woman's right to have an abortion under specific circumstances. Because this is a count variable with a fixed total, it is most natural to work with a logit link and binomial error function. The predictors that we used are the year of measurement (1=1983; 2=1984; 3=1985; 4=1986) and religion (1=Roman Catholic, 2=Protestant; 3=Other; 4=No religion). The effect of year of measurement is assumed to be class-dependent and the effect of religion is assumed to be the same for all classes.

We estimated models with 1 to 5 classes, and the 4-class model turned out to performs best in terms of the BIC criterion. We also estimated more restricted models in which the time effect is assumed to be linear and/or the time effect is assumed to be class independent. These models did not describe the data as well as our four-class model, which indicates that the time trend is non linear and heterogeneous.

[INSERT TABLE 11 ABOUT HERE]

The parameters obtained with the 4-class model appear in Table 11. The parameter means across classes indicate that the attitudes are most positive at the last time point and most negative at the second time point. Furthermore, the effects of religion show that people without religion are most in favor and Roman Catholics and Others are most against abortion. Protestants have a position that is close to the no-religion group.

The class-specific parameters indicate that the 4 latent classes have very different intercepts and time patterns. The largest class 1 is most against abortion and class 3 is most in favor of abortion. Both latent classes are very stable over time. The overall level of latent class 2 is somewhat higher than of class 1, and it shows somewhat more change of the attitude over time. People belonging to latent class 4 are very instable: at the first two time points they are similar to class 2, at the third time point to class 4, and at the last time point again to class 2 (this can be seen by combining the intercepts with the time effects). Class 4 could therefore be labeled as random responders. It is interesting to note that in a three-class solution the random-responder class and class two are combined. Thus, by going from a three- to a four-class solution one identifies the interesting group with less stable attitudes.

Vermunt and Van Dijk (2001) used the same empirical example to illustrate the similarity between LC regression models and random-coefficients, multilevel, or hierarchical models. Using terminology from multilevel modeling, the time variable is a level-1 predictor and religion a level-2 predictor. The effect of the level-1 predictor time is allowed to vary across level 2 units, in this case individuals. The LC regression output can be transformed into the usual output produced by a standard multilevel or hierarchical model -- means, variances, covariances of the intercept and the three time effects -- by elementary statistical operations. The most important part of this multilevel output is what appears in the last two columns of Table 11.

A difference between LC regression analysis and standard hierarchical models is that the former does not make strong assumptions about the distribution of the random coefficients. LC regression models can, therefore, be seen as non-parametric hierarchical models in which the distribution of the random coefficients is approximated by a limited number of mass points (= latent classes). As shown by Vermunt and Van Dijk (2001), the LC approach has the practical advantage of being much less computationally intensive than parametric models, and substantively, easier-to-interpret results are often obtained.

<u>Example: application to choice-based conjoint studies</u>

The LC regression model is a popular tool for the analysis of data from conjoint experiments in which individuals rate or choose between sets of products having different attributes (Wedel and

Kamakura, 1998). The objective is to determine the effect of product characteristics on the rating or the choice probabilities. LC analysis is used to identify subgroups, or market segments, for which these effects differ.

For illustration of LC analysis of data obtained from choice-based conjoint experiments, we use a generated data set. The products are 10 pairs of shoes that differ on 3 attributes: Fashion (0=traditional, 1= modern), Quality (0=low, 1=high), and Price (ranging from 1 to 5). Eight choice sets offer 3 of the 10 possible alternative products to 400 individuals. Each choice task consists of indicating which of the three alternatives they would purchase, with the response "none of the above" allowed as a fourth choice option.

The model that is used is a multinomial logit model with choice-specific predictors, also referred to as the conditional logit model. Let *J* be the number of choice sets, *M* the number of choices per set, and *Q* the number of predictors. A particular set, choice, and predictor is denoted by *j*, *m*, and *q*, respectively. The regression model of interest is

$$P(y_j = m \mid X = c) = \frac{\exp\left(\sum_{q=1}^{Q} \beta_{qc} w_{jmq}\right)}{\sum_{m'=1}^{M} \exp\left(\sum_{q=1}^{Q} \beta_{qc} w_{jm'q}\right)}$$

Here, $P(y_j = m \mid X = c)$ denotes the probability that someone belonging to class *c* selects choice-alternative *m* in choice set *j*. The predictors we use are the three product attributes (fashion, quality, and price), as well as a dummy variable for the "none" category.

[INSERT TABLE 12 ABOUT HERE]

The BIC values indicated that the three-class model is the model that should be preferred. The parameter estimates obtained with the 3-class model are reported in Table 12. As can be seen, FASHION has a major influence on choice for class 1, QUALITY for class 2, and both FASHION and QUALITY affect the choice for class 3. The price effect is similar for all three classes. The Wald test for the equality of effects between classes indicates that the difference in price effects across classes is not significant. The price effects could, therefore, be assumed to be class independent.

[INSERT TABLE 13 ABOUT HERE]

In addition to the conditional logit model, which shows how the predictors affect the likelihood of choosing one alternative over another, differentially for each class, we specified a second logit model to describe the latent class variable as a function of the covariates sex and age. Table 13 shows that females turn out to belong more often to class 1 and males to class 3. Younger persons have a higher probability of belonging to class 1 (emphasize Fashion in choices) and older persons are most likely to belong to class 2 (emphasize Quality in choices).

In conclusion, the LC regression model offers computational and interpretive advantages over the more traditional hierarchical modeling approach that tends to overfit data (Andrews, Ansari, and Currim, 2002). In our example, we used the BIC criteria to select a parsimonious number of classes. However, researchers who prefer the results to show *higher* levels of individual variation in regression coefficients can obtain such with LC regression models by simply increasing the number of latent classes to produce the desired amount of variation.

4.3 LC Analysis as an Alternative to K-means Clustering

An important application of LC analysis is clustering (Banfield and Raftery, 1993; McLachlan and Peel, 2000; Vermunt and Magidson, 2002). Actually, we already saw several cluster-like applications. The traditional LC model was used to construct a typology of survey respondents using a set of categorical indicators. We also showed that simple mixture models like mixtures of normals or mixtures of Poisson distributions can be used for clustering purposes.

In this section, we will concentrate on LC analysis as a tool for cluster analysis with *continuous* indicators. These LC models can be seen as multivariate extensions of the mixtures of univariate normals discussed above. Instead of assuming a univariate normal distribution, we assume multivariate normal distributions within latent classes. The most general form of the mixture model concerned assumes that each latent class has its own set of means, variances, and covariances. More formally,

$$f(\boldsymbol{y}|\vartheta) = \sum_{c=1}^{C} P(X = c) f(\boldsymbol{y}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) .$$

Here, $\boldsymbol{\mu}_c$ denotes the vector with class-specific means and $\boldsymbol{\Sigma}_c$ the class-specific variance-covariance matrix. Note that, contrary to traditional LC modeling, it is not necessary to assume local independence between the indicators.

The above LC cluster model is similar to the model used in discriminant analysis. An important difference is, of course, that in cluster analysis group membership is unobserved or latent, which is reason that LC cluster analysis is sometimes referred to as latent discriminant analysis.

[INSERT FIGURE 5 ABOUT HERE]

The first part of Figure 5 depicts a data set that we will use to illustrate the LC cluster model for continuous variables. Three measures are available to diagnose diabetes: Glucose, Insulin, and SSPG (steady-state plasma glucose) (see Fraley and Raftery, 1998). In addition to these measures, we have information on the clinical diagnosis consisting of the three categories "normal", "chemical diabetes", and "overt diabetes". However, in practice, a gold standard is not available in cluster applications. Our objective here is to construct a mixture model that yields a classification that is close to the clinical diagnosis, without use of the information on the clinical diagnosis. We use this data set to demonstrate the flexibility of LC clustering compared to other clustering methods. The gold standard makes it possible to judge whether the methods do what we want them to do.

LC cluster analysis is a model-based clustering procedure. As such it is a probabilistic and more flexible alternative to K-means clustering. K-means clustering performs well under very strict conditions; that is, if indicators are locally independent and if error variances are cluster invariant and equal across indicators ($\mathbf{\Sigma}_c = \sigma^2 \mathbf{I}$). These implicit assumptions of K-means imply that in a 3-dimensional scatter plot each cluster has the form of a sphere with the same radius and in each 2-dimensional plot, each cluster will have the form of a sphere with the same radius. The assumption of equal error variances across indicators is the reason that in K-means clustering it is advised to standardize the variables prior the analysis. While standardization often improves the situation, it does not solve the problem because equating the variance in the total sample is not the same as equating the within-group variances (Magidson and Vermunt, 2002).

Having a closer look at Figure 5, it can easily be seen that it is impossible to describe the shape of the three diabetes clusters by a K-means model; that is, by 3 spheres with the same radius. The within-cluster variances are very different across clusters and across indicators. Moreover, the glucose and insulin indicators are strongly correlated within the group with overt diabetes. Nevertheless, since the clusters are well separated, a reliable cluster method should be able to yield a three-cluster solution that is similar to the clinical classification.

[INSERT TABLE 14 ABOUT HERE]

The problems associated with K-means are confirmed by the test results reported in Table 14. We estimated 1- to 5-cluster models, each with four different specifications of the variance-covariance matrix: diagonal (=local independence) and equal across classes, diagonal and unequal, glucose-insulin covariance and unequal, and all covariances and unequal. It can be seen that when the specifications are too restrictive, one needs 5 and 4 clusters, respectively. Actually, with the first K-means-like specification, even more than 5 clusters are needed.

Although the BIC values indicate that the two additional local dependencies (y_1-y_3 and y_2-y_3) in the full model are not needed (compare the three-cluster solutions for the last two specifications), the fit measures also show that both the model with the fully unrestricted covariance matrix and the model with only the glucose-insulin covariance detect the correct three-cluster solution. This means that working with a model with insufficient restrictions does not harm in this example, but this is not always the case.

The middle part of Figure 5 shows the five nearly spherical clusters identified with the most restricted specification we used. Similar results would have been obtained with K-means. The lower part of Figure 5 depicts the 3-cluster solution that turned out to be the best according to the BIC criterion. It can be seen, that the 3 clusters identified by this model are very similar to the clinical classification. Our 3-cluster solution is smoother in the sense that some of the overlap between the clinical classes disappears, which is, of course, what can be expected from a statistical model. The correspondence between the 3-cluster and the clinical classification is 87%, which is only slightly lower than the 93% correct classifications of a quadratic discriminant analysis (in which cluster membership is treated as known).

The LC cluster model cannot only be applied with continuous indicators, but also with indicators of other scale types and different combinations of scale types. Depending on the scale type, one will specify the most appropriate within-cluster distribution for the indicator concerned. This yields a general cluster model for mixed-mode data (Hunt and Jorgensen, 1999; Vermunt and Magidson, 2002). Note that the traditional LC model is the special case in which all indicators are categorical variables.

4.4 Other Developments in LC Modeling

In this chapter, we presented what we believe to be the most important types of LC models. We did not discuss LC models for specific types of data, such as longitudinal, event history, or multilevel data. Some of these models are mixture regression models and can, therefore, be handled within the LC regression framework (Vermunt, 1997, 2007). Another important class of models for longitudinal data are latent or hidden Markov models which can be used to study how individuals move across latent classes over time (see, e.g., Collins and Lanza, 2010; Langeheine and Van de Pol, 1994; Vermunt, Tran, and Magidson, 2008). Moreover, for data with a multilevel structure, Vermunt (2003) proposed a variant of the LC cluster model yielding a clustering of both higher- and lower-level units.

We presented LC models that can be used for scaling. There also exist more sophisticated LC scaling models, which can be obtained by imposing certain constraints on the parameters of the traditional LC model. Examples are LC models for probabilistic Guttman scaling, LC models with order constraints, LC Rasch models, LC models for preference data, and LC models for distance data (see, Heinen, 1996; Dayton 1998; Böckenholt, 2002; Croon, 2002).

Another more advanced type of LC model we would like to mention is the Lisrel-type framework for categorical variables developed by Hagenaars (1990) and extended by Vermunt (1997). Any type of LC models with categorical indicators, including LC models for transition data and sophisticated LC scaling models, are special cases of this general model. A limitation if this approach is that it is restricted to categorical indicators.

A final class of models we would like to mention are more sophisticated restricted mixtures of multivariate normals than those discussed above. LC models have been proposed in which the class-specific covariance matrices are constrained by means of principal component (Fraley and Raftery, 1998) or factor-analytic (Yung, 1997) structures, or by structural equation models (Jedidi, Jagpal, and DeSarbo, 1997).

REFERENCES

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

Andrews, R.L., Ansari, A., and Currim, I.S. (2002). Hierarchical bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery, *Journal of Marketing Research*, 39, 87-98.

Bakk, Zs., Tekle, F.B., and Vermunt, J.K. (2013). Estimating the association between latent class membership and external variables using bias adjusted three-step approaches. *Sociological Methodology*, 43, 272-311.

Bakk, Zs., and Vermunt, J.K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling*, 23, 20-31.

Banfield, J.D., and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803-821.

Böckenholt, U. (2002). Comparison and choice: analyzing discrete preference data by latent class scaling models. J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis*. Cambridge University Press.

Clogg, C.C., and Goodman, L.A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, 79, 762-771.

Collins, L.M., and Lanza, S.T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New Jersey: John Wiley & Sons

Croon, M.A. (2002) Ordering the classes. J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis*. Cambridge University Press.

Dayton, C.M., and Macready, G.B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83, 173-178.

Dayton, C.M. (1998). *Latent Class Scaling Models*. Thousand Oakes: Sage Publications.

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.

Everitt, B.S., and Hand, D.J. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

Fraley, C. and Raftery, A.E. (1998). *How many clusters? Which clustering method? Answers via model-based cluster analysis*. Department of Statistics, University of Washington: Technical Report no. 329.:

Formann, A.K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87, 476-486.

Goodman, L.A. (1974a). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.

Goodman, L.A. (1974b). 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. (1988). Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods and Research*, 16, 379-405.

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

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

Hunt, L., and Jorgensen. M. (1999). Mixture model clustering using the MULTIMIX program. *Australian and New Zeeland Journal of Statistics*, 41:153-172.

Jedidi, K., Jagpal, H.S., and DeSarbo, W.S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16, 39-59.

Landis, J.R., and Koch, G.G. (1977). The measurement of observer agreement for categorical data, *Biometrics*, 33, 159-174.

Langeheine, R., and Van de Pol, F. (1994). Discrete-time mixed Markov latent class models. A. Dale and R.B. Davies (eds.), *Analyzing social and political change: a casebook of methods*, 171-197.

Langeheine, R., Pannekoek, J., Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, 24, 492-516.

Lazarsfeld, P.F., and Henry, N.W. (1968). *Latent Structure Analysis*. Boston: Houghton Mill.

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

Magidson J. and Vermunt, J.K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, 20, 37-44.

McCutcheon, A.L. (1987). *Latent Class Analysis*, Sage University Paper. Newbury Park: Sage Publications.

McGrath, K., and Waterton, J. (1986). *British Social Attitudes, 1983-1986 Panel Survey*. London: Social and Community Planning Research, Technical Report.

McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Rindskopf, R., and Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine*, 5, 21-27.

Uebersax J.S. (1999) Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models. *Applied Psychological Measurement*, 23, 283-297.

Uebersax J.S., and Grove, W.M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9, 559-572.

Van der Heijden P.G.M., Gilula, Z., and Van der Ark, L.A. (1999). On a relationship between joint correspondence analysis and latent class analysis. *Sociological Methodology*, 29, 147-186.

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

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, 213-239.

Vermunt, J.K. (2007). Growth models for categorical response variables: standard, latent-class, and hybrid approaches. In: K. van Montfort, H. Oud, and A. Satorra (eds.), *Longitudinal Models in the Behavioral and Related Sciences*, 139-158. Mahwah, NJ: Erlbaum.

Vermunt, J.K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18, 450-469.

Vermunt, J.K., and Magidson, J. (2000). *Latent GOLD User's Guide*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., and Magidson, J. (2002). Latent class cluster analysis. J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis*, 89-106. Cambridge University Press.

Vermunt, J.K., and Magidson, J. (2013). *Latent GOLD 5.0 Upgrade Manual*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., Tran, B., and Magidson, J. (2008). Latent class models in longitudinal research. In: S. Menard (ed.), *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, pp. 373-385. Burlington, MA: Elsevier.

Vermunt, J.K. and Van Dijk. L. (2001). A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter*, 13, 6-13.

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.

Yung, Y.F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, 62, 297-330.

Table 1: Results from Various LC Models Fit to the GSS'82 Data

Model		BIC_{LL}	L^2	*df*	p value	% Reduction in L^2(H_0)
<u>Sample of white respondents</u>						
Traditional						
H_0	1-class	5787.0	257.3	29	2.0×10^{-38}	0.0%
H_{1C}	2-class	5658.9	79.5	22	2.0×10^{-8}	69.1%
H_{2C}	3-class	5651.1	22.1	15	0.11	91.4%
H_{3C}	4-class	5685.3	6.6	8	0.58	97.4%
Nontraditional						
H_{1C+}	2-class + {y_3-y_4} direct effect	5606.1	12.6	20	0.89	95.1%
H_{2F}	Basic 2-factor	5640.1	11.1	15	0.75	95.7%
<u>Sample of black respondents</u>						
Traditional						
H'_0	1-class	2402.1	112.1	29	1.0×10^{-11}	0.0%
H'_1	2-class	2389.6	56.9	22	.00006	49.2%
H'_{2C}	3-class	2393.8	18.3	15	.25	83.7%
H'_{3C}	4-class	2427.6	9.4	8	.31	91.6%
Nontraditional						
H'_{1C+}	2-class + { y_3-y_4} direct effect	2360.2	15.2	20	.77	86.4%
H'_{2F}	Basic 2-factor	2387.0	11.5	15	.72	89.7%
<u>Full sample (multiple-group analysis)</u>						
Traditional						
M_0	1-class	8185.1	400.0	64	4.3×10^{-50}	0%
M_1	2-class	8013.8	169.5	56	2.4×10^{-13}	57.6%
M_{2C}	3-class unrestricted (comp. heterogeneity)	8077.4	40.4	30	.10	89.9%
M_{2CR}	3-class restricted (partial homogeneity)	7953.0	49.4	48	.42	87.7%
M_{2CRR}	3-class restricted (comp. homogeneity)	7962.1	73.3	50	.02	81.7%
M_{3CR}	4-class restricted (partial homogeneity)	7989.8	27.0	40	.94	93.3%
Nontraditional						
M_{2F}	basic 2-factor unrestricted	8059.6	22.6	30	.83	94.4%
M_{2FR}	basic 2-factor restricted	7934.9	31.3	48	.97	92.2%

Table 2: Parameter Estimates for the 3-Class LC Model by Sample

	White Sample			Black Sample		
	Class 1 Ideal	Class 2 Believers	Class 3 Skeptics	Class 1 Ideal	Class 2 Believers	Class 3 Skeptics
LC Probabilities	0.62	0.2	0.18	0.49	0.33	0.18
Conditional Probabilities						
(A) PURPOSE						
Good	0.89	0.92	0.16	0.87	0.91	0.19
Depends	0.05	0.07	0.22	0.08	0.04	0.17
Waste	0.06	0.01	0.62	0.05	0.05	0.65
(B) ACCURACY						
Mostly True	0.61	0.65	0.04	0.54	0.65	0.01
Not True	0.39	0.35	0.96	0.46	0.35	0.99
(C) UNDERSTANDING						
Good	1	0.32	0.75	0.95	0.37	0.68
Fair, poor	0	0.68	0.25	0.05	0.63	0.32
(D) COOPERATION						
Interested	0.95	0.69	0.64	0.98	0.56	0.64
Cooperative	0.05	0.26	0.26	0.01	0.37	0.25
Impatient/ Hostile	0	0.05	0.1	0	0.07	0.11

Table 3: Descriptive Information and Parameter Estimates from 3-Class and 2-Factor LC Models Obtained with the Landis and Koch Data

	Descriptive Information			3 Classes			2 Factors (joint probabilities)			
	% of slides rated positive	% of ratings that agree with					Factor1=1 (true -)		Factor1=2 (true +)	
		5+ raters	6+ raters	Class 1	Class 2	Class 3	Factor2=1 (– bias)	Factor2= 2 (+ bias)	Factor2= 1 (– bias)	Factor2= 1 (+ bias)
Class size				0.44	0.37	0.18	0.36	0.19	0.30	0.16
Rater 6	21%	64%	58%	**0.47**	0.00	0.00	0.00	0.01	**0.23**	0.86
Rater 4	27%	70%	62%	**0.59**	0.00	0.06	0.00	0.05	**0.37**	0.92
Rater 3	38%	80%	64%	0.85	0.00	0.01	0.00	0.00	0.83	0.83
Rater 1	56%	82%	64%	1.00	0.06	0.51	0.06	**0.47**	0.99	1.00
Rater 7	56%	85%	66%	1.00	0.00	0.63	0.01	**0.58**	0.99	1.00
Rater 5	60%	80%	64%	1.00	0.06	0.76	0.06	**0.72**	0.99	1.00
Rater 2	67%	75%	61%	0.98	0.15	0.99	0.13	**0.99**	0.97	1.00

Table 4: Results from Various LC Models Fit to Landis and Koch Data

Model		BIC_{LL}	L^2	Bootstrap p value	% Reduction in L^2(H_0)
Traditional					
H_0	1-class	1082.3	476.8	.00	0.0%
H_1	2-class	707.9	64.2	.00	86.5%
H_{2C}	3-class	699.6	17.7	.49	96.3%
H_{3C}	4-class	729.4	9.3	.79	98.0%
Nontraditional					
H_{2C+}	3-class + {y_4y_5} direct effect	698.0	11.3	.83	97.6%
H_{2FR}	Restricted[7] basic 2-factor	688.4	11.3	.90	97.6%

Table 5: Values for Bivariate Residuals obtained Under Various Models for the Sample of White Respondents

2-way table	Model					
	H_0	H_1	H_{2C}	H_{3C}	H_{2C+}	H_{2F}
$\{y_1y_2\}$	61.6	0.1	0.1	0.0	0.0	0.0
$\{y_1y_3\}$	0.5	0.7	0.1	0.0	0.2	0.0
$\{y_1y_4\}$	10.6	0.0	0.1	0.0	0.2	0.1
$\{y_2y_3\}$	0.3	1.1	0.0	0.0	0.0	0.0
$\{y_2y_4\}$	8.6	0.4	0.3	0.2	0.2	0.4
$\{y_3y_4\}$	43.4	32.3	2.4	0.0	0.0	0.2

Table 6: Bivariate Residuals Obtained Under Various Models for Landis and Koch data

2-way table*	Model					
	Traditional			Non-traditional		
	H_0	H_1	H_{2C}	H_{2C+}	H_{2FR}	H_{2FRC}
$\{y_2y_5\}$	66.4	8.4	0.0	0.0	0.0	0.1
$\{y_4y_6\}$	38.0	7.2	4.5	0.0	0.0	0.0
$\{y_2y_7\}$	66.7	5.2	0.0	0.0	0.1	0.1
$\{y_5y_7\}$	77.2	3.3	0.1	0.1	0.2	0.2
$\{y_1y_2\}$	54.5	1.7	0.1	0.0	0.1	0.1
$\{y_3y_6\}$	28.0	1.3	0.0	0.0	0.0	0.0
$\{y_3y_5\}$	47.7	1.1	0.1	0.1	0.2	0.1
$\{y_4y_5\}$	24.5	0.0	0.7	0.6	0.6	1.2

* These are the 2-way tables for which the bivariate residuals were larger than 1 under any of the reported models (other than H_0)

Table 7: Parameter Estimates for the 3-Class LC model of Partial Homogeneity (Model M_{2CR}) and the Corresponding LC 2-Factor Model M_{2FR}

	3 Classes			2 Factors (marginal probabilities)			
	Class 1	Class 2	Class 3	Factor 1		Factor 1	
	Ideal	Believers	Skeptics	Level 1	Level 2	Level 1	Level 2
LC Probabilities							
Whites	0.68	0.15	0.17	0.81	0.19	0.85	0.16
Blacks	0.51	0.30	0.19	0.79	0.21	0.70	0.31
Conditional Probabilities							
PURPOSE							
Good	0.89	0.90	0.16	0.90	0.20	0.76	0.78
Depends	0.06	0.06	0.21	0.06	0.21	0.09	0.07
Waste	0.05	0.04	0.63	0.05	0.59	0.15	0.15
ACCURACY							
Mostly True	0.60	0.64	0.01	0.63	0.02	0.50	0.55
Not True	0.40	0.36	0.99	0.37	0.98	0.50	0.45
UNDERSTANDING							
Good	0.94	0.32	0.74	0.79	0.76	0.92	0.26
Fair, poor	0.06	0.68	0.26	0.21	0.24	0.08	0.74
COOPERATION							
Interested	0.95	0.57	0.65	0.86	0.66	0.90	0.50
Cooperative	0.05	0.35	0.25	0.12	0.24	0.09	0.38
Impatient/ Hostile	0.00	0.08	0.10	0.02	0.10	0.01	0.12

Table 8: Parameter Estimates for the 2-Factor Restricted Multi-Group LC Model with Covariates

	Factor	
	F_1	F_2
Covariates (gammas)		
G: Group		
WHITE	0	-0.20
BLACK	0	0.20
E:Years of Education		
<8	0	2.19
8-10	0	0.97
11	0	0.08
12	0	-0.34
13-15	0	-1.01
16-20	0	-1.89
Indicator Variables (lambdas)		
A: PURPOSE	2.26	0
B: ACCURACY		
mostly true	-1.34	0
not true	1.34	0
C:UNDERSTANDING		
Good	0	-5.14
Fair/Poor	0	5.14
D: COOPERATION	0.98	1.26

Table 9: Test Results for Generated Mixture of Normals Data

Model	Log-Likelihood	BIC_{LL}	Number of Parameters
Equal variances			
1-Class	-2177.75	4369.31	2
2-Class	-2066.99	4161.61	4
3-Class	-2050.78	4143.00	6
4-Class	-2046.25	4147.75	8
Unequal Variances			
2-Class	-2048.14	4130.81	5
3-Class	-2047.78	4150.83	8
4-Class	-2045.41	4166.80	11

Table 10. Observed and Estimated Frequency Distribution of Packs of Hard-Candy Purchased During Last 7 Days under the 1-Class and 3-Class Poisson Model

Number of	Frequencies		
Packages	Observed	1-Class Model	3-Class Model
0	102	8.43	101.67
1	54	33.63	54.63
2	49	67.11	50.03
3	62	89.28	53.89
4	44	89.09	47.25
5	25	71.11	34.14
6	26	47.30	22.00
7	15	26.97	14.37
8	15	13.46	11.02
9	10	5.97	10.18
10	10	2.38	10.17
11	10	0.86	9.97
12	10	0.29	9.20
13	3	0.09	7.90
14	3	0.03	6.32
15	5	0.01	4.72
16	5	0.00	3.30
17	4	0.00	2.18
18	1	0.00	1.36
19	2	0.00	0.80
20	1	0.00	0.45

Table 11: Parameter Estimates for the Abortion Example

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

Table 12: Parameter Estimates for Conditional Logit Model in Conjoint Study Example

	Class1	Class2	Class3	Wald for no effect	Wald for equal effects
FASHION	3.03	-0.17	1.20	494.74	216.37
QUALITY	-0.09	2.72	1.12	277.96	171.16
PRICE	-0.39	-0.36	-0.56	144.48	3.58
NONE	1.29	0.19	-0.43	82.39	59.26

Table 13: Parameter Estimates for the Latent Variable Regression for Conjoint Study Example

	Class1	Class2	Class3	Wald
Intercept	0.37	0.00	-0.37	8.22
SEX				
Male	-0.66	-0.34	1.01	24.15
Female	0.66	0.34	-1.01	
AGE				
16-24	1.02	-0.15	-0.87	62.76
25-39	-0.59	-0.37	0.96	
40+	-0.43	0.52	-0.09	

Table 14: Test Results for Diabetes Data

Model	Log-Likelihood	BIC_{LL}	Number of Parameters
Equal and diagonal			
1-Cluster	-2750.13	5530.13	6
2-Cluster	-2559.88	5169.52	10
3-Cluster	-2464.78	4999.24	14
4-Cluster	-2424.46	4938.49	18
5-Cluster	-2392.56	**4894.60**	22
Unequal and diagonal			
1-Cluster	-2750.13	5530.13	6
2-Cluster	-2446.12	4956.94	13
3-Cluster	-2366.92	4833.38	20
4-Cluster	-2335.38	**4805.13**	27
5-Cluster	-2323.13	4815.47	34
Unequal and full			
1-Cluster	-2546.83	5138.46	9
2-Cluster	-2359.12	4812.80	19
3-Cluster	-2308.64	**4761.61**	29
4-Cluster	-2298.13	4790.34	39
5-Cluster	-2284.97	4813.79	49
Unequal and y_1-y_2 free			
1-Cluster	-2560.40	5155.64	7
2-Cluster	-2380.27	4835.19	15
3-Cluster	-2320.57	**4755.61**	23
4-Cluster	-2303.14	4760.56	31
5-Cluster	-2295.05	4784.19	39

Figure 1: Barycentric Coordinate Display for 3-Class Model

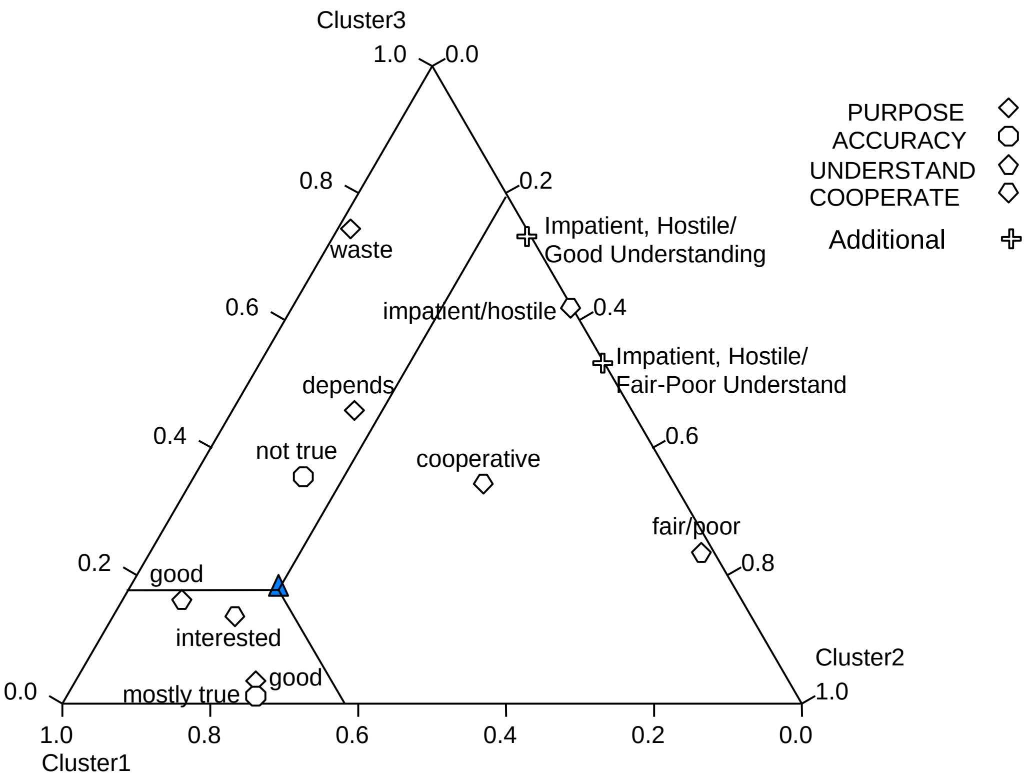

Figure 2: Bi-plot for 2-Factor Model with Covariates

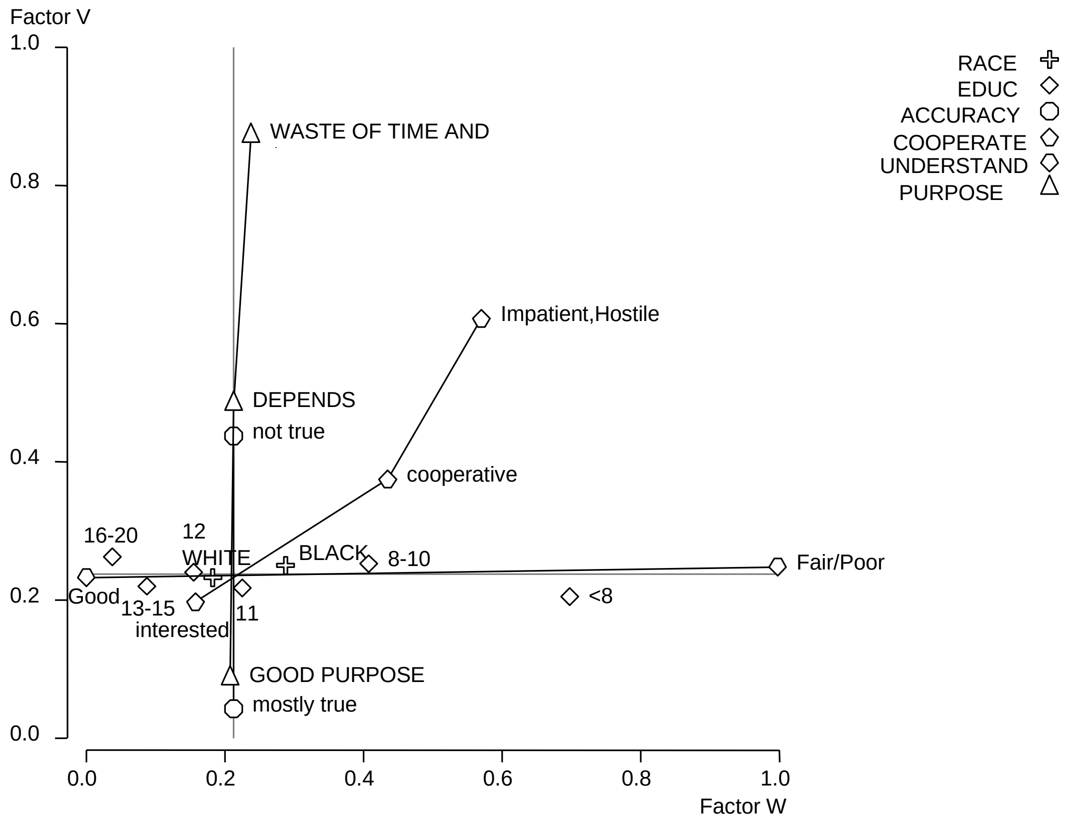

Figure 3: Simulated Distribution from a Mixture of Two Normals

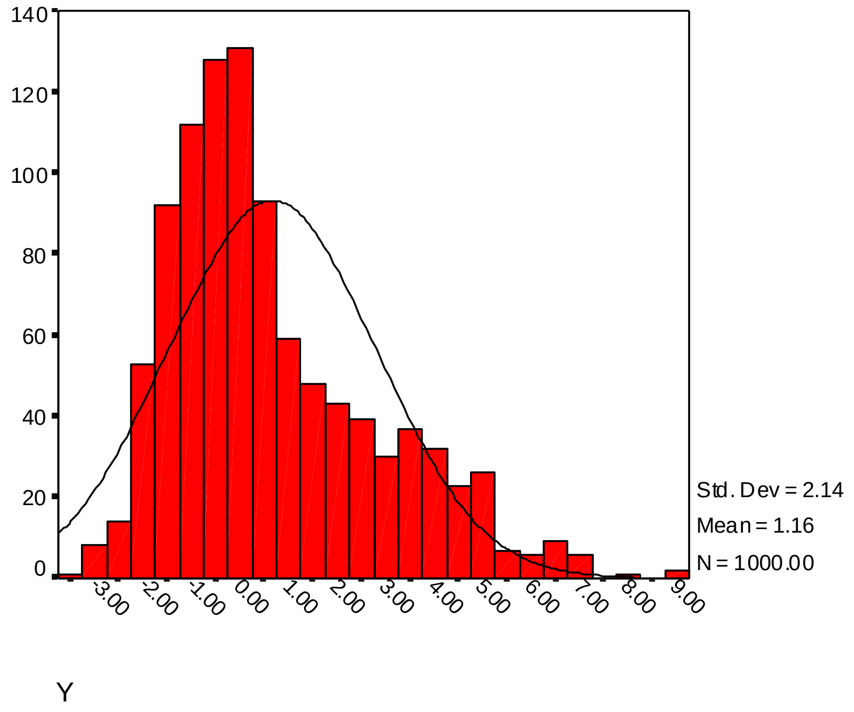

Figure 4: Simulated 2-Class LC Regression Model

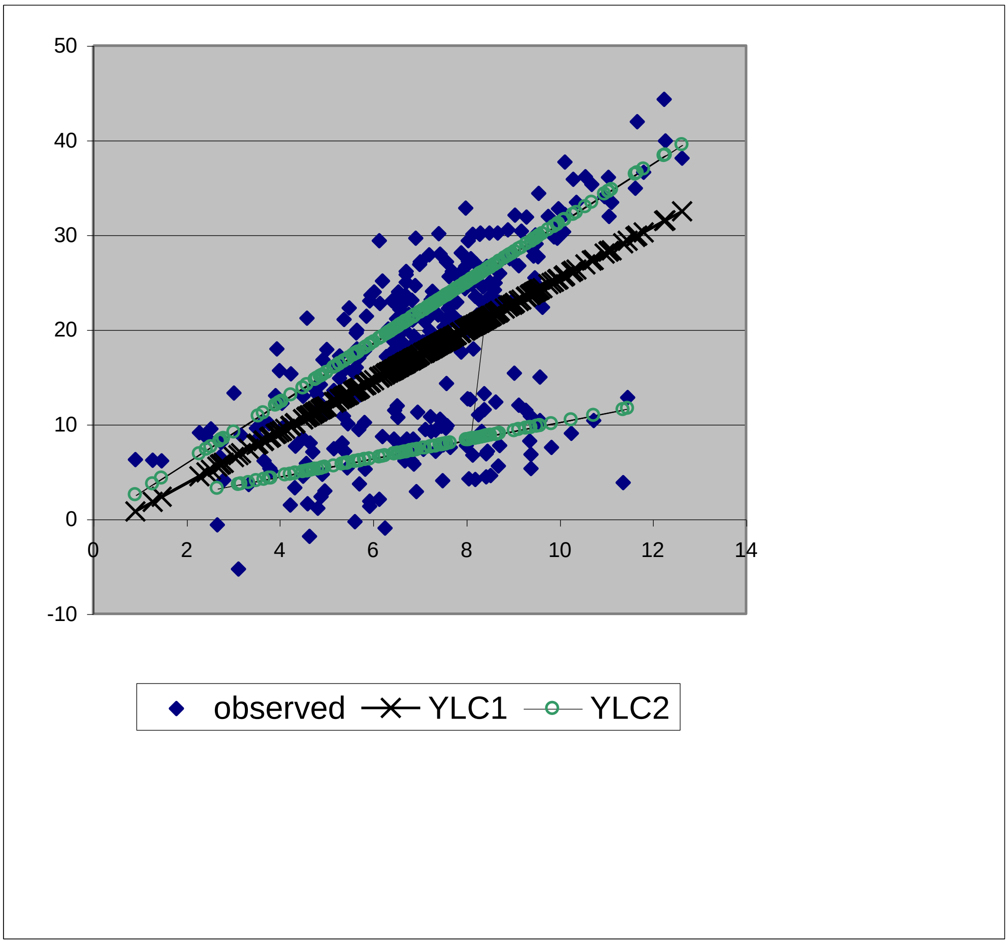

Figure 5: Matrix Scatter Plot of Diabetes Data set for the Clinical Classification, the K-Means-Like 5-Cluster Solution, and the Final 3-Cluster Solution