LATENT CLASS FACTOR AND CLUSTER MODELS, BI-PLOTS AND RELATED GRAPHICAL DISPLAYS

Jay Magidson and Jeroen K. Vermunt

Statistical Innovations and Tilburg University

jay@statisticalinnovations.com and j.k.vermunt@kub.nl

The authors wish to thank Jeremy F. Magland, Leo A Goodman and Peter G.M. van der Heijden for helpful comments.

We propose an alternative method of conducting exploratory latent class analysis that utilizes latent class factor models, and compare it to the more traditional approach based on latent class cluster models. We show that when formulated in terms of R mutually independent, dichotomous latent factors, the LC factor model has the same number of distinct parameters as an LC cluster model with R+1 clusters. Analyses over several data sets suggest that LC factor models typically fit data better and provide results that are easier to interpret than the corresponding LC cluster models. We also introduce a new graphical "bi-plot" display for LC factor models and compare it to similar plots used in correspondence analysis and to a barycentric coordinate display for LC cluster models. We conclude by describing various model extensions and an approach for eliminating boundary solutions that we have implemented in a new computer program called Latent GOLD®.

Key Words: Latent class models, biplots, factor analysis, cluster analysis, correspondence analysis

1. INTRODUCTION

Latent class (LC) analysis is becoming one of the standard data analysis tools in social, biomedical, and marketing research. While the traditional LC model described by Lazarsfeld and Henry (1968) and Goodman (1974a, 1974b) contains only nominal indicator variables, variants have been proposed for ordinal (Clogg 1988; Uebersax 1993; Heinen 1996) and continuous indicators (Wolfe 1970; McLachlan and Basford 1988; Fraley and Raftery 1998), as well as for combinations of variables of different scale types (Lawrence and Krzanowski 1996; Moustaki 1996; Hunt and Jorgensen 1999; Vermunt and Magidson 2001). This paper concentrates on exploratory LC analysis with nominal and ordinal indicators.

In an exploratory LC analysis, the usual approach is to begin by fitting a 1-class (independence) model to the data, followed by a 2-class model, a 3-class model, etc., and continuing until a model is found that provides an adequate fit (Goodman 1974a, 1974b; McCutcheon 1987). We refer to such models as LC cluster models since the T nominal categories of the latent variable serve the same function as the T clusters desired in cluster analysis (McLachlan and Basford 1988; Hunt and Jorgensen 1999; Vermunt and Magidson 2001).

Van der Ark and Van der Heijden (1998) and Van der Heijden, Gilula and Van der Ark (1999) showed that exploratory LC analysis can be used to determine the number of dimensions underlying the responses on a set of nominal items. A LC model with three classes, for example, can be seen as a two-dimensional model similar to a two-dimensional joint correspondence analysis (JCA). However, within the context of LC analysis, a more natural manner of specifying the existence of two underlying dimensions for a set of items is to specify a model containing two latent variables.

Goodman (1974b), Haberman (1979), and Hagenaars (1990, 1993) proposed restricted 4-class LC models yielding confirmatory LC models with two latent variables. Their approach is confirmatory since, as in confirmatory factor analysis, it requires a priori knowledge on which items are related to which latent variables. In *exploratory* data analysis settings, we do not know beforehand which items load on the same latent variable.

Hence, in exploratory analyses with several latent variables, this approach has limited practical applicability.

In this paper, we propose combining the exploratory model fitting strategy of the traditional latent class model with the possibility of increasing the number of latent variables to study the dimensionality of a set of items. Our alternative model fitting sequence involves increasing the number of latent variables (factors) rather than the number of classes (clusters). We call the latter sequence the LC factor approach because of the natural analogy to standard factor analysis. The basic LC factor model contains R mutually independent, dichotomous latent variables. To exclude higher-order interactions, logit models are specified on the response probabilities. An interesting feature of the basic R-factor model is that it has exactly the same number of parameters as an LC cluster model with T = R+1 clusters. In section 2, we describe the two types of exploratory LC models using the log-linear formulation introduced by Haberman (1979).

Section 3 compares the use of LC cluster and factor models and describes various graphical displays that facilitate the interpretation of the results obtained from these models. Specifically, we consider some variations of the ternary diagram originally proposed by Van der Ark and Van der Heijden (1998) for LC cluster models, and introduce a new display (called a "bi-plot") for LC factor models to represent various kinds of information in a 2-dimensional factor space. These two graphs are compared to each other and to similar displays used in correspondence analysis.

Section 4 presents some final remarks regarding the applicability of these models. For a more complete version of this paper see Magidson and Vermunt (2001).

2. TWO APPROACHES FOR EXPLORATORY LATENT CLASS ANALYSIS

In this section we describe and compare two competing alternative approaches for exploratory LC analysis. The traditional approach utilizes LC cluster models, while the alternative is based on LC factor models. For the sake of simplicity of exposition, below we use the log-linear formulation of LC models introduced by Haberman (1979). In

Appendix A, we give the alternative probability formulation of the two types of LC models, as well as the relationship between the two formulations.

2.1 The Latent Class Cluster Model

For concreteness, consider 4 nominal variables denoted A, B, C, and D. Let X represent a nominal latent variable with T categories. The log-linear representation of the LC cluster model with T classes is:

$$\ln(F_{ijklt}) = \lambda + \lambda_t^X + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{it}^{AX} + \lambda_{jt}^{BX} + \lambda_{kt}^{CX} + \lambda_{lt}^{DX} \qquad (1)$$

where i = 1,2,…,I; j=1,2,…,J; k=1,2,…K; l=1,2,…L; and t=1,2,…T.

For convenience in counting distinct parameters and without loss of generality, we choose the following "dummy coding" restrictions to identify the parameters[1]:

$$\lambda_1^X = \lambda_1^A = \lambda_1^B = \lambda_1^C = \lambda_1^D = 0$$

$\lambda_{i1}^{AX} = \lambda_{j1}^{BX} = \lambda_{k1}^{CX} = \lambda_{l1}^{DX} = 0$ for i = 1,2,…,I; j=1,2,…,J; k=1,2,…K; l=1,2,…L;

and $\lambda_{1t}^{AX} = \lambda_{1t}^{BX} = \lambda_{1t}^{CX} = \lambda_{1t}^{DX} = 0$ for t = 2,3,…,T.

As can be seen, the LC model described in equation (1) has the form of a log-linear model for the five-way frequency table cross-classifying the 4 observed variables and the latent variable; that is, the table with cell entries F_{ijklt}. The assumed model contains one-variable terms ("main effects") associated with the latent variable X and the four observed indicators A, B, C, and D, as well as all two-variable "interaction" terms that involve X which pertain to the association between X and each of the observed indicators. The one-variable effects are included because we do not wish to impose constraints on the univariate

[1] See Haberman (1979) for an alternative set of identifying restrictions based on ANOVA effects coding.

marginal distributions. The assumption that the observed responses to A, B, C, and D are mutually independent given X = t ("local independence") is imposed by the omission of all interaction terms pertaining to the associations between the indicators. As shown in Appendix A, this set of conditional independence assumptions can also be formulated in another way, yielding the probability formulation for the LC model.

Note that for the 1-class model, since T=1, the model described in equation (1) reduces to the usual log-linear model of mutual independence between the 4 observed variables:

$$\ln(F_{ijkl}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D \,. \qquad (2)$$

More generally, for models with any number of variables, we will denote the model of mutual independence as H_0, and use it as a baseline to assess the improvement in fit to the data of various LC models. The number of distinct parameters[2] in the model of independence as described in equation (2) is:

$$\text{NPAR(indep)} = (\text{I-1}) + (\text{J-1}) + (\text{K-1}) + (\text{L-1})$$

Expressing the number of distinct parameters in the model described in equation (1) as a function of NPAR(indep), yields:

$$\begin{aligned}\text{NPAR(T)} &= (\text{T-1}) + \text{NPAR(indep)} \text{ x } [1 + (\text{T-1})] \\ &= (\text{T-1}) + \text{NPAR(indep)} \text{ x T}\end{aligned}$$

The number of degrees of freedom (DF) associated with the test of model fit is directly related to the number of distinct parameters in the model tested[3].

$$\begin{aligned}\text{DF(T)} &= \text{IJKL} - \text{NPAR(T)} - 1 \\ &= \text{IJKL} - [1 + \text{NPAR(indep)}] \text{ x T}\end{aligned}$$

[2] By convention, we do not count λ as a distinct parameter because of the redundancy to the overall sample size, and we subtract 1 from the number of cells when computing degrees of freedom.

[3] It is customary when one or more distinct parameters are unidentified or not estimable (a boundary solution), to adjust the DF, increasing it by the number of such unidentified or not estimable parameters.

Beginning with this baseline model (T=1), each time the number of latent classes (T) is incremented by 1 the number of distinct parameters increases by 1 + NPAR(indep), and, as a consequence, the degrees of freedom are reduced by 1 + NPAR(indep). The first additional parameter is the main effect for the additional latent class, and the NPAR(indep) further parameters correspond to the effects of each observed (manifest) variable on this additional latent class.

2.2 The Latent Class Factor Model

Certain LC models can be interpreted in terms of 2 or more component latent variables by treating those components as a joint variable (Goodman 1974b; McCutcheon 1987; Hagenaars 1990). For example, a 4-category latent variable X = {1, 2, 3, 4} can be re-expressed in terms of 2 dichotomous latent variables V = {1,2} and W = {1, 2} using the following correspondence:

	W=1	W=2
V=1	X =1	X = 2
V=2	X =3	X = 4

Thus, X=1 corresponds with V=1 and W=1, X=2 with V=1 and W=2, X=3 with V=2 and W=1, and X=4 with V=2 and W=2.

The LC cluster model given in (1) with T = 4 classes can be re-parameterized as an *unrestricted* LC factor model with two dichotomous latent variables V and W as follows:

$$\begin{aligned}\ln(F_{ijklrs}) = {} & \lambda + \lambda_r^V + \lambda_s^W + \lambda_{rs}^{VW} + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ir}^{AV} + \lambda_{jr}^{BV} + \lambda_{kr}^{CV} + \lambda_{lr}^{DV} \\ & + \lambda_{is}^{AW} + \lambda_{js}^{BW} + \lambda_{ks}^{CW} + \lambda_{ls}^{DW} + \lambda_{irs}^{AVW} + \lambda_{jrs}^{BVW} + \lambda_{krs}^{CVW} + \lambda_{lrs}^{DVW} ,\end{aligned} \qquad (3)$$

The correspondence between the two representations is that the one-variable terms pertaining to X are now written as $\lambda_{2(r-1)+s}^{X} = \lambda_r^V + \lambda_s^W + \lambda_{rs}^{VW}$, and the two-variable terms involving X as $\lambda_{i,2(r-1)+s}^{AX} = \lambda_{ir}^{AV} + \lambda_{is}^{AW} + \lambda_{irs}^{AVW}$, $\lambda_{j,2(r-1)+s}^{BX} = \lambda_{jr}^{BV} + \lambda_{js}^{BW} + \lambda_{jrs}^{BVW}$, etc. It is easy to verify that this re-parameterization does not alter the *number* of distinct parameters in the model.

We define the *basic* R-factor LC model as a *restricted* factor model that contains R mutually independent, dichotomous latent variables, containing parameters ("factor loadings") that measure the association of each latent variable on each indicator. Specifically, the basic R-factor model is defined by placing two sets of restrictions on the unrestricted LC factor model. The resulting 2-factor LC model is a restricted form of the 4-class LC cluster model. Without these restrictions, the 2-factor model would be unconstrained and would be equivalent to a 4-cluster model.

The first set of restrictions sets to zero each of the 3-way and higher-order interaction terms. For the basic 2-factor model, we have $\lambda_{irs}^{AVW} = \lambda_{irs}^{BVW} = \lambda_{irs}^{CVW} = \lambda_{irs}^{DVW} = 0$. After imposing these restrictions, the 2-variable terms in the basic 2-factor model become

$$\lambda_{i,2(r-1)+s}^{AX} = \lambda_{ir}^{AV} + \lambda_{is}^{AW}, \qquad \lambda_{j,2(r-1)+s}^{BX} = \lambda_{jr}^{BV} + \lambda_{js}^{BW}, \qquad \text{etc.}$$

For variable A, λ_{ir}^{AV} represents the loading of A on factor V and λ_{is}^{AW} denotes the loading of A on factor W, etc. By fixing the three-variable terms to be equal to zero, we obtain a model that is conceptually similar to standard exploratory factor analysis: each of the factors may have an effect on each indicator, but there are no higher-order interaction terms. Constraints of this form are necessary to allow the four latent classes to be expressed as a cross-tabulation of two latent variables and thus are essential for distinguishing the LC factor model from the LC cluster model.

The second set of restrictions imposes mutual independence between the latent variables. For the 2-factor model, this latter restriction imposes independence in the 2-way table <VW>. This restriction makes the model more similar to standard *exploratory* factor analysis. We relax this assumption in section 4, when we present *confirmatory* LC factor models.

Although the basic R-factor model is a special case of an LC *cluster* model containing 2^R classes, we show in Appendix A that because of the restrictions of the type given above, the basic R-factor LC model is actually comparable to an LC cluster model with only T = R+1 clusters in terms of parsimony. This large reduction in number of parameters will be sufficient to achieve model identification in many situations. That is, in practice, it will frequently be the case that the basic R-factor will be identified when the LC cluster model with 2^R classes is not.

TABLE 1
Equivalency Relationship between LC Cluster and Basic LC Factor Models
(Example with 5 Dichotomous Variables)

LC Cluster Models				Basic LC Factor Models		
Number of Latent Classes	Number of Parameters	Degrees of Freedom		Number of Factors	Number of Parameters	Degrees of Freedom
1	5	26		0	5	26
2	11	20		1	11	20
3	17	14		2	17	14
4	23	8		3	23	8
5	29	2		4	29	2

Table 1 verifies the equivalence in number of parameters (and the associated degrees of freedom) between the various identified LC cluster models and the corresponding basic LC factor models in the case of 5 dichotomous indicator variables. From this table we can also calculate, for example, that the basic LC 2-factor model requires 23 – 17 = 6 fewer parameters than the 4-class LC cluster model. This reduction corresponds to the 5 restrictions $\lambda_{irs}^{AVW} = \lambda_{irs}^{BVW} = \lambda_{irs}^{CVW} = \lambda_{irs}^{DVW} = \lambda_{irs}^{EVW} = 0$, plus the restriction that V and W are independent.

We conclude this section by noting an important difference between our LC factor model and the LC models with several latent variables proposed by Goodman (1974b), Haberman (1979), McCutcheon (1987), and Hagenaars (1990, 1993). The basic LC factor model described above includes all factor loadings between the latent variables and the indicators. This means that no assumptions need be made about which indicators are related to which latent variables. This makes this LC factor model better suited for exploratory data analysis than the LC models with several latent variables described in the literature.

Thus far we have described two alternative approaches for exploratory LC analysis, one involving the fitting of LC cluster models, the other fitting basic LC factor models. In the next section we consider some examples to illustrate and compare their performance on real data and introduce graphical displays that facilitate the interpretation of the obtained results.

3. EXAMPLES AND GRAPHICAL DISPLAYS

Comparison of the two approaches for exploratory LC analysis across several data sets found that the factor approach resulted in a more parsimonious and easier to interpret model almost every time. Since our selection of data sets was not random, we do not present those results here. Rather, for purposes of illustration, this section considers the analysis from two data sets where a basic 2-factor model fits the data. In the first example, the comparable cluster model also provides an acceptable (but not as good) fit to the data; in the second example, the comparable cluster model provides a *much* worse fit, one that is not acceptable for these data.

3.1. Example 1: 1982 General Social Survey Data

Our first example, taken from McCutcheon (1987) and reanalyzed by Van der Heijden, Gilula, and Van der Ark (1999) involves four categorical variables from the 1982 General Social Survey. Two items are evaluations of surveys by white respondents and the other two are evaluations of these respondents by the interviewer. A summary of various LC models fit to these data is given in Table 2.

TABLE 2: Results from Various LC Models Fit to General Social Survey Data

Model	**Model Description**	**BIC**	**L^2**	**DF**	**p-value**	**% Reduction in L^2(H_0)**
H_0	1-class	51.6	257.26	29	$2.0\text{x}10^{-38}$	0 %
H_1	2-class	-76.7	79.34	22	$2.1\text{x}10^{-8}$	69.1%
H_{2C}	3-class	-98.7	21.89	$15+2^{\dagger}$	0.19	91.5%
H_{2F}	Basic 2-factor	-109.6	10.93	$15+2^{\dagger}$	0.86	95.7%
H_3	4-class	-72.0	6.04	$8+3^{\dagger}$	0.87	97.7%
H_{R2F}	Restricted 2-factor	-140.9	22.17	$22+1^{\dagger}$	0.51	91.4%

H_{1F3}	1-factor (3 levels)	-71.7	77.25	21	$2.3x10^{-8}$	70.0%

† DF is increased by these boundary solutions

Model H_0 is the baseline model given in equation (2) which specifies mutual independence between all four variables. Model H_0 is a 1-class LC model (a 1-cluster model) which can also be interpreted as the equivalent 0-factor LC model. Since L^2 = 257.26 with DF = 29, this model is rejected. Next, consider the 2-class model (H_1) that can be interpreted as either a 2-cluster model or the equivalent 1-factor model where the factor is dichotomous. The L^2 is now reduced to 79.34, a 69.1% reduction from the baseline model, but too high to be acceptable with DF = 22.

Next, consider the two 15-DF models[4] -- H_{2C}, the 3-cluster model and H_{2F}, the basic 2-factor model. Each of these models provide an adequate fit to the data, although the factor model fits better, the L^2 being half that of the comparable cluster model. For comparison, Table 3 also provides results for the 4-cluster model (H_3). Among the first 5 models listed in Table 3, H_{2F} is preferred according to the BIC criteria. The last 2 models in Table 3 are extended models that will be discussed in the next section.

TABLE 3
Comparison of results from the 3-Cluster Model with the Basic 2-Factor Model
Conditional Membership Probability of being in Cluster j =1,2,3 (for Model H_{2C})
or level 1 of Factor k=1,2 (for Model H_{2F})

	Model H_{2C}			Model H_{2F}	
	Cluster 1	Cluster 2	Cluster 3	Factor1(1)	Factor2(1)
Indicators					
PURPOSE					
Good	0.72	0.25	0.03	0.83	0.71
Depends	0.38	0.17	0.45	0.65	0.28
Waste	0.24	0.02	0.73	0.59	0 †
ACCURACY					
Mostly True	0.73	0.26	0.01	0.83	0.83
Not True	0.50	0.15	0.35	0.71	0.28

[4] For both models H_{2C} and H_{2F}, the maximum likelihood solution contains 2 boundary solutions and hence, by convention (see note 3) we increased the DF by 2. Adding the number of parameters estimated on the boundary to the number of degrees of freedom is a convention in LC analysis (see, for instance, McCutcheon, 1987). In our opinion, there is no good reason to do so, but it is outside the scope of this paper to present alternative testing methods for situations in which boundary estimates occur. For model H_{2C}, McCutcheon (1987) reported an adjusted DF of 16, increasing the usual DF by only 1 because the solution reported was not fully converged and contained, therefore, only 1 boundary solution. The solution presented in Van der Heijden et. al. (1999) is the same solution as that presented here (containing 2 boundary solutions) but they also misreport the DF to be 16 instead of 17.

UNDERSTAND					
good	0.76	0.08	0.16	0.89	0.53
Fair, poor	0 †	0.77	0.23	0.28	0.71
COOPERATE					
Interested	0.70	0.17	0.13	0.86	0.58
Cooperative	0.27	0.40	0.33	0.38	0.51
Impatient/ Hostile	0 †	0.39	0.61	0 †	0.35
Overall Probability	0.62	0.21	0.17	0.78	0.57

† indicates a boundary solution

Table 3 compares results obtained from the 3-cluster Model (H_{2C}) with that from the basic 2-factor model (H_{2F}). The cell entries in the left-most columns are "rescaled parameter estimates" suggested by Van der Heijden, Gilula, and Van der Ark (1999), and represent the estimated *conditional* probabilities of being a member of one of the three clusters. The right-most columns contain corresponding quantities for the basic 2-factor model, representing the estimated probabilities of being at level 1 for each of the 2 factors. *Unconditional* membership probabilities for the clusters and for level 1 of the factors are given in the last row of the table.

Graphical displays of the conditional probabilities reported in Table 3 are useful in comparing results between the two models. For the 3-cluster model H_2, Van der Heijden, Gilula, and Van der Ark (1999, Figure 4) present a ternary diagram for visualizing the results and show the close relationship to 2-dimensional plots produced by joint correspondence analysis (JCA). A slightly modified graphic, referred to here as a "barycentric coordinate" display is given in Figure 1 for the 3-cluster model H_{2C}. The shaded triangle in Figure 1 with lines emanating to the sides represents the overall sample which is plotted at the point corresponding to the unconditional membership probabilities for the clusters.

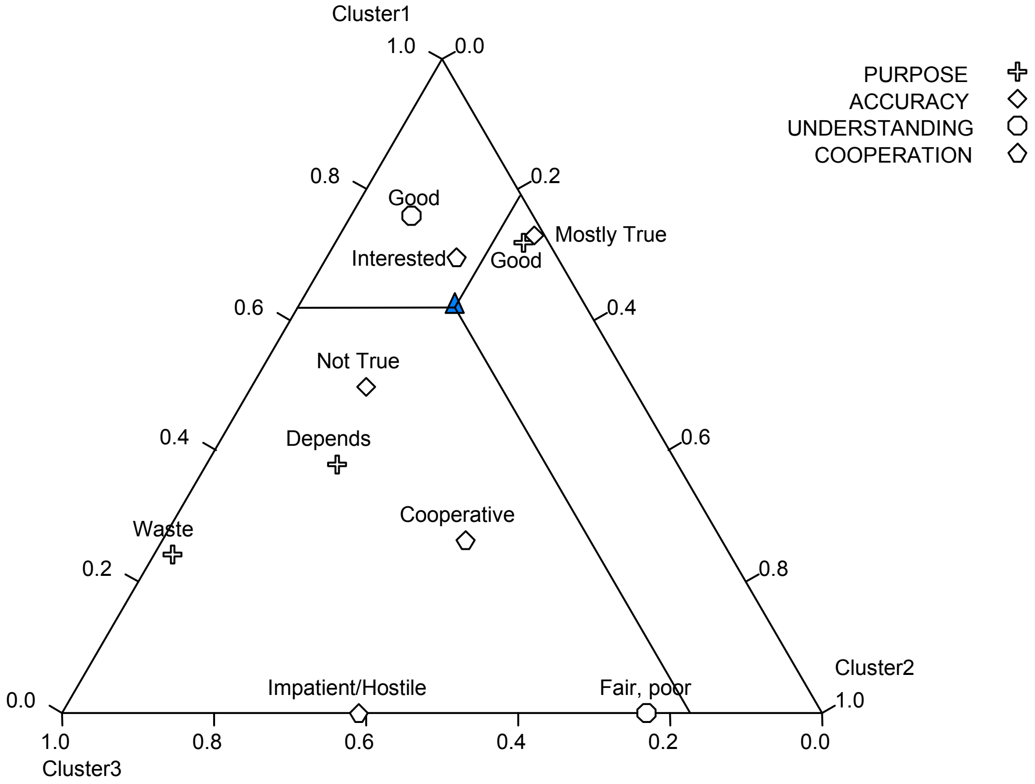

FIGURE 1. Barycentric Coordinate Display of Results for Model H_{2C}

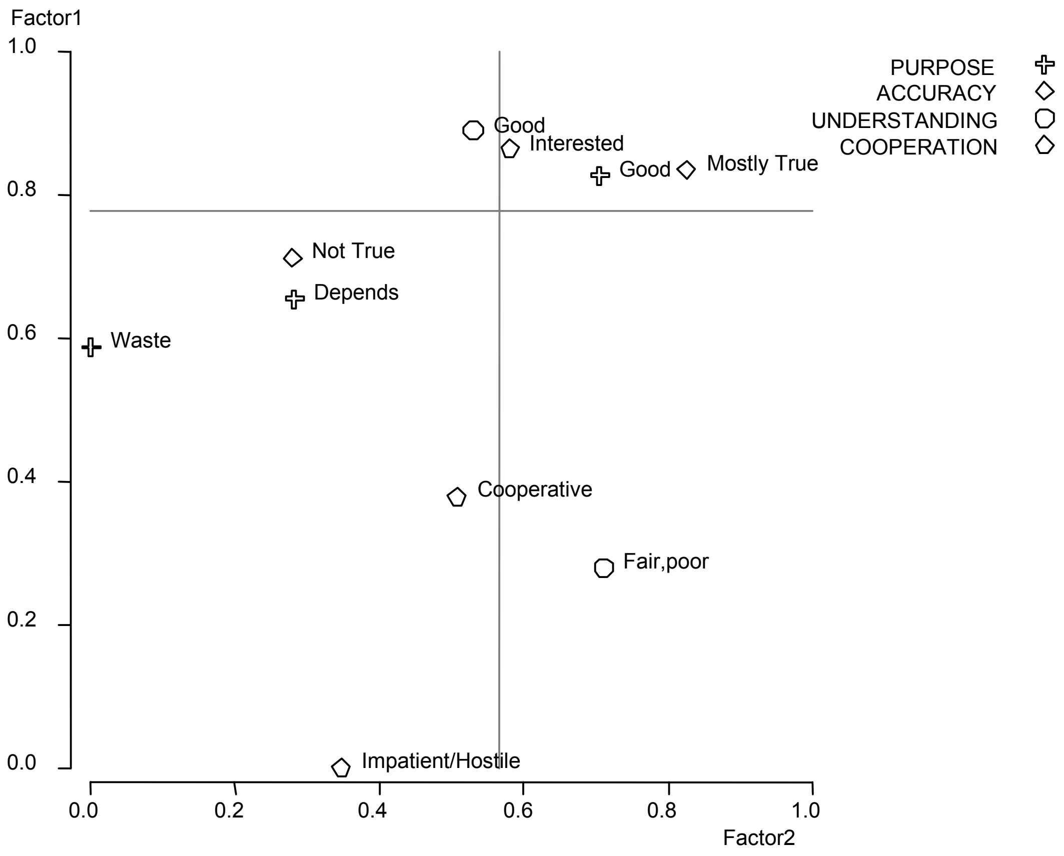

FIGURE 2. Bi-plot of Results Reported for Model H_{2F}

A different display for LC factor-models called the "bi-plot"[5] (Vermunt and Magidson, 2000) is given in Figure 2 for the 2-factor model H_{2F}. For comparability to the barycentric coordinate plot where cluster 1 is assigned to the top vertex, we take factor 1 to be the *vertical* axis and factor 2 the horizontal. By comparing these plots we can see the large degree of similarity between the models, the primary difference being the relative positioning of COOPERATION = Impatient/ Hostile and UNDERSTANDING = Fair, poor.

[5] In the context of correspondence analysis, the term "biplot" refers to a particular joint display of points representing both the rows and columns of a frequency table (Greenacre, 1993). On the other hand, Gower and Hand (1996) stress that the "bi" in biplots arises from the fact that cases and variables are presented in the same plots. In Vermunt and Magidson (2000), we chose the term "bi-plot" because of the similarity of our plots to the plots used in correspondence analysis. However, despite the fact that in most of our examples we depict only variable categories, it is also possible to depict cases (or answer patterns) in our plots as we illustrated in our Figures 4, 6 and 8. For more detail about our plots see Appendix B.

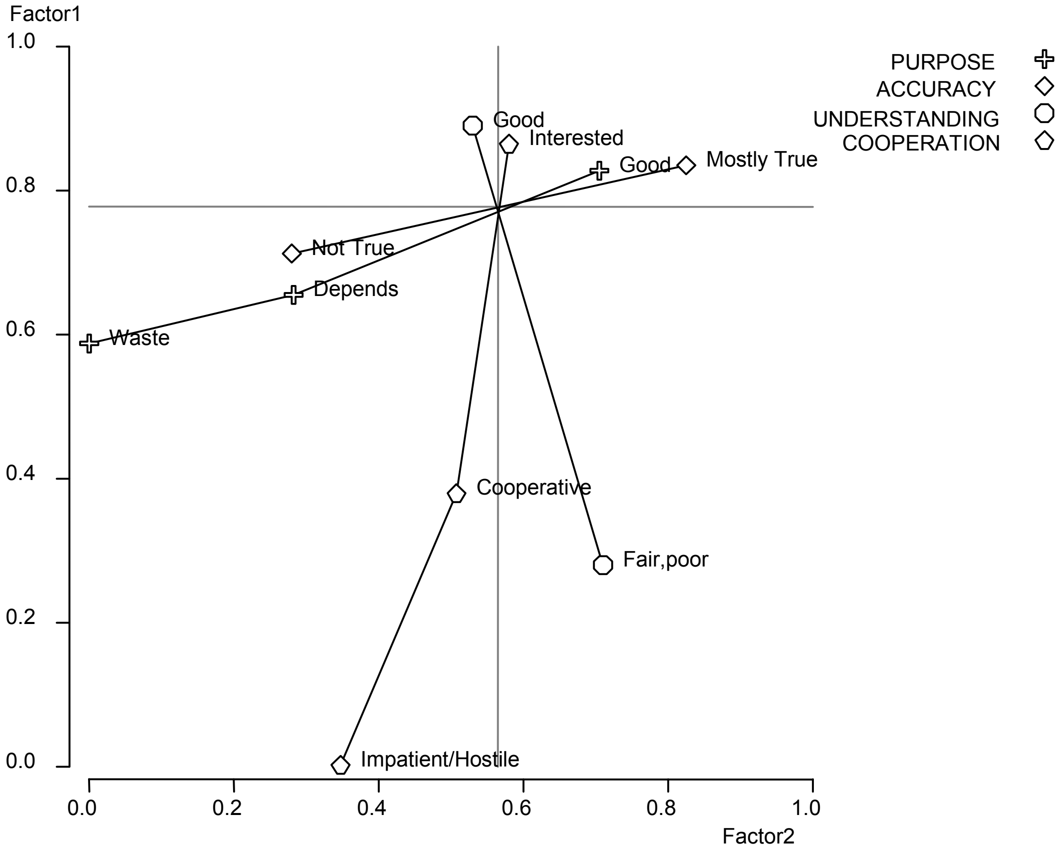

FIGURE 3. Bi-plot for Model H_{2F} with Lines connecting categories of a Variable

Lines connecting the categories of a variable can make it easier to see to which factor the variables are most related. For example, Figure 3 shows that separation between the categories of the two respondent evaluation variables, PURPOSE and ACCURACY occurs primarily along Factor 2 (the horizontal axis in Figure 3) while for the two interviewer evaluation variables, UNDERSTANDING and COOPERATION separation occurs primarily along Factor 1 (the vertical axis). This makes clear that Factor 1 pertains primarily to the interviewer valuation while Factor 2 pertains primarily to the respondent valuation. These two factors are not only distinct (i.e., the 1-factor model H_1 does not fit these data) but according to model H_{2F}, they are mutually independent.

Since our models yield estimated membership probabilities for each individual case, both displays can easily be extended to include points for individual cases and covariate levels as well as any other desired groupings of the cases (see Appendix B). Our methodology is unified in the sense that the same methods and models that yield our displays for LC cluster models also yield the bi-plots for the LC factor models. Our barycentric coordinate display can be more easily extended in this manner than the methods proposed by Van der Heijden, Gilula, and Van der Ark (1999) with the ternary diagram. In our next example we will illustrate the inclusion in our plots of cases by including specific cases with selected response patterns. Then in section 4, we show how the display of *all* response patterns can be used to identify a natural ordering between the classes (when such an ordering exists), and we describe two different approaches for overlaying covariate values (levels) onto the displays.

The bi-plots offer several advantages over the related plots produced in correspondence analysis (CA) even when the data justifies a 2-dimensional CA solution[6]. That is because the 2-dimensional CA solution is closely related to the 3-cluster solution (Gilula and Haberman 1986; De Leeuw and Van der Heijden, 1991) which we have found typically does not fit the data as well as the 2-factor solution. As suggested in this paper, the LC factor models generally provide simpler explanations of data than LC

[6] An extensive comparison between the LC cluster model and (joint) correspondence analysis is given by Van der Heijden, Gilula and Van der Ark (1999). They showed that (joint) correspondence analysis is very similar to what we labeled the LC cluster model. More precisely, a 2-dimensional joint correspondence

cluster models and the related canonical models used in CA and principal components analysis.

Our LC factor model is more closely related to traditional factor analysis than to CA. Advantages over traditional factor analysis include 1) the variables can include different scale types – nominal, ordinal, continuous and/or counts, 2) solutions are typically uniquely identified and interpretable without the need for a rotation – there is no rotational indeterminacy, and 3) factor scores can be obtained for each case without the need for additional assumptions. Like traditional factor analysis, LC factor analysis can be used as a first step in a more confirmatory analysis. In the next section we describe a more confirmatory analysis of the data analyzed above.

4. SOME EXTENSIONS OF THE BASIC LC FACTOR MODEL

In this section we consider some modifications and extensions of the basic LC factor model that may be of interest in certain situations. First, although in example 1 we treated the trichotomous variables COOPERATE (A) and PURPOSE (C) as nominal, they can be treated as ordinal in several different ways. The most straight-forward approach is to assume the middle category to be equidistant from the others and modify the log-linear model described in equation (3) by using the uniform scores v_i^A and v_k^C

v_i^A = {0 if i = 1, 0.5 if i=2, 1 if i = 3}

v_k^C = {0 if k = 1, 0.5 if k=2, 1 if k = 3}

for the categories of variables A and C. Secondly, analogous to confirmatory factor analysis, we may wish to allow the two factors V and W to be correlated (with association parameter γ_{rs}^{VW}) and restrict the variables COOPERATION (A) and UNDERSTANDING (B) to load only on factor 1 and PURPOSE (C) and ACCURACY (D) to load only on factor 2. The log-linear representation for a confirmatory model of this type as compared to the basic 2-factor model in Appendix A is as follows:

analysis can describe exactly the results – the estimated frequencies in all two-way tables -- of a 3-cluster model.

$$\gamma_{rs}^{VW} \neq 0;$$
$$\lambda_{ir}^{AV} = \lambda_{r}^{AV} v_{i}^{A};\ \lambda_{ks}^{CW} = \lambda_{s}^{CW} v_{k}^{C}; \quad \text{where} \quad i,k = 1,2,3;\ j,l,r,s = 1,2;$$
$$\lambda_{is}^{AW} = \lambda_{js}^{BW} = \lambda_{jr}^{CV} = \lambda_{ks}^{DV} = 0.$$

The results of fitting this restricted 2-factor model (H_{R2F}) are reported in Table 3. These suggest that this model fits the data very well (L^2 = 22.17, DF=23; p = .51). The corresponding bi-plot is shown in Figure 4 below.

FIGURE 4. Bi-plot for Model H_{R2F} with Lines connecting the categories of a Variable

Our examples thus far utilized only dichotomous factors. To extend the factor model so that any factor may contain more than 2 ordered levels, we assign equidistant numeric scores between 0 and 1 to the levels of the factor. Clogg (1988) and Heinen (1996) used the same strategy for defining LC models that are similar to certain latent trait models. The use of fixed scores for the factor levels in the various two-way interaction terms guarantees that each factor captures a single dimension. For factors with more than two levels, in the bi-plot we display conditional means rather than conditional probabilities (see Appendix B). Note that if we assign the score of 0 to the first level and 1 to the last level (or vice versa), for dichotomous factors the conditional mean equals the conditional probability of being at level 2 (or level 1).

Finally, the extension to include covariates in a log-linear LC model is straightforward. To illustrate the use of covariates and the extension to a 3-level factor, we will use the depression scale data for white respondents from the “Problems of Everyday Life” study conducted in 1972 by Pearlin (Pearlin and Johnson 1977) as reported separately for males and females (Schaeffer,1988). Persons who reported having the symptom during the previous week were coded 1, all others 0. The symptoms measured were lack of enthusiasm, low energy, sleeping problem, poor appetite and feeling hopeless.

Gender was included in the model as an *active* covariate (see the discussion in Appendix B on ‘active vs. inactive covariates’). Note that in the case of a single covariate, the log-linear LC model is identical whether GENDER is treated as a covariate or as another indicator (Clogg 1981; Hagenaars 1990).

TABLE 4
Results from Various LC Models Fit to the Depression Data

Model	**Model**					**% Reduction**
	Description	**BIC**	**L^2**	**DF**	**p-value**	**in L^2(H_0)**
H_0	1-class	672.8	1097.1	57	2.3×10^{-192}	0
H_1	2-class	-233.7	138.5	50	3.1×10^{-10}	87.4%
H_{2C}	3-class	-260.5	59.6	43	0.05	94.6%
H_{2F}	Basic 2-factor	-274.6	45.5	43+1†	0.37	95.9%
H_{1F3}	1-factor (3-levels)	-297.8	67.0	49	0.05	93.9%

† df is increased by these boundary solutions

Table 5 Conditional Probabilities Estimated under the 3-Cluster model and the 1-Factor 3-level model

	3-Cluster Model			1-Factor 3-level Model		
	Cluster1	Cluster2	Cluster3	Level1	Level2	Level3
Cluster Size	0.46	0.44	0.10	0.45	0.45	0.10
ENTHUS						
Lack of enthusiasm	0.26	0.82	0.96	0.26	0.81	0.98
No	0.74	0.18	0.04	0.74	0.19	0.02
ENERGY						
Low energy	0.03	0.63	0.95	0.03	0.61	0.99
No	0.97	0.37	0.05	0.97	0.39	0.01
SLEEP						
Sleeping problem	0.10	0.37	0.78	0.09	0.38	0.79
No	0.90	0.63	0.22	0.91	0.62	0.21
APPETITE						
poor appetite	0.04	0.22	0.73	0.04	0.24	0.72
No	0.96	0.78	0.27	0.96	0.76	0.28
HOPELESS						
Hopeless	0.03	0.10	0.67	0.02	0.13	0.61
No	0.97	0.90	0.33	0.98	0.87	0.39

Table 4 shows the results from fitting various LC models to these data. The traditional strategy required 3 classes as neither the 1- or 2-class models provided adequate solutions. We see again that the basic 2-factor model fits the data better than the comparable 3-cluster model. The results for the 3-cluster solution are shown in Table 5 in terms of conditional response probabilities. Notice that those probabilities conditional on cluster 2 are ordered between the corresponding probabilities conditional on clusters 1 and 3, a pattern that is consistent with the depression scale being uni-dimensional, and suggests that we consider fitting a 3-level 1-factor model to these data.

Table 5 shows that the 3-level factor solution is very similar to that given by the 3-class solution. Both suggest that 10% of the population are in the “depressed” group (cluster 3 in the cluster model and level 3 in the factor model), and the rest are about equally distributed among the “healthy” (cluster 1) and the “troubled” cluster 2. The 3-level model provides an acceptable fit to these data and only contains one parameter more than the 2-class model (see Table 5). Unlike the 3-class extension to the 2-class model which requires 7 additional parameters, the 3-level model provides an attractive alternative to the 3- (unordered) class model. The BIC suggests that the 3-level 1-factor model should be preferred over all models including the basic 2-factor model.

In our experience with various data, increasing the number of levels in a factor does often provide a significant improvement in fit. This is, however, not always the case. For example, with our first data set we found that 2 distinct factors were required to provide an adequate fit to the data. In that situation, increasing the number of levels from 2 to 3 in the single factor solution provides no benefit. Table 3 shows only a slight, non-significant reduction in the L^2 due to the inclusion of the additional parameter -- from 79.34 for the 1-factor 2-level solution to 77.25 for the 1-factor 3-level solution. On the other hand, in the present example, the addition of this single parameter causes a reduction of the L^2 from 138.5 for the 1-factor 2-level solution to 67.0 under the 1-factor 3-level model (see Table 5).

An informative graph can provide an attractive alternative to a table (such as TABLE 5) when the goal is to determine whether a natural ordering exists among a set of clusters. For example, a standard profile plot will show immediately that the conditional probabilities associated with cluster 2 always fall between the corresponding conditional probabilities associated with clusters 1 and 3.

FIGURE 5: Barycentric Coordinate Display of the 64 Response Patterns for Males and Females based on the 3-class Model (H_{2c})

.

Note: The area of each triangle is proportional to the estimated expected frequency associated with the corresponding response pattern (subject to a minimum size).

As an alternative to the profile plot, we will now examine the implications obtained from a barycentric coordinate display (FIGURE 5) of the 3-cluster solution which includes a point for each observation (i.e., each observed response pattern). Note the obvious pattern that the points appear primarily along the left and right sides of the triangle, and not along the base. This visual pattern can be interpreted as follows -- among persons who are likely to be “troubled“ (those with response patterns plotted near the top vertex, associated with cluster 2), there is a substantial amount of overlap with the other clusters. Some of these cases also have a substantial probability of belonging to the “healthy” cluster and some have a substantial probability of belonging to the “depressed” cluster. However, there is virtually *no* overlap between those likely to be “healthy” and those likely to be “depressed” (the inner part of the base of the triangle contains no points). This asymmetric pattern suggests that cluster 2 (“troubled”) is the middle cluster.

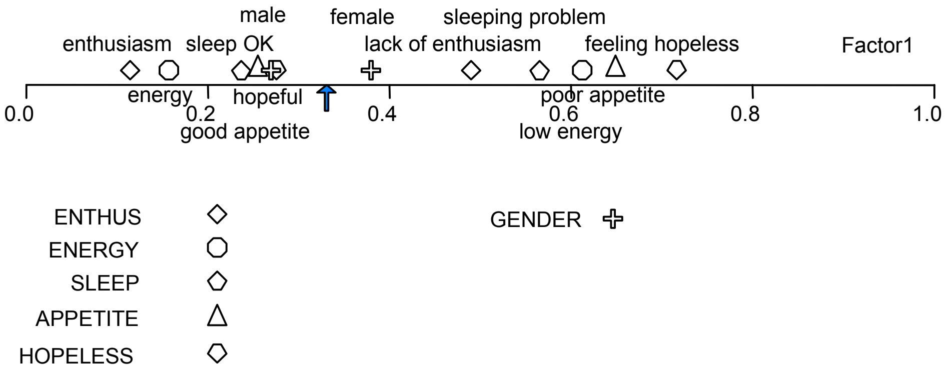

FIGURE 6: One-dimensional plot associated with the 3-level Factor Model

In both the 3-cluster model and the 3-level 1-factor model, we find that GENDER has a significant relationship with the latent variable, females being more likely to be in the depressed group. Figure 6 displays a 1-dimensional plot resulting from the 3-level factor model (the bi-plot reduces to one dimension in the case of a single factor). In general, inclusion of covariates in a model can provide useful descriptive information on the latent variable(s).

5. FINAL REMARKS

This paper presented a new method for performing exploratory LC analysis. Rather than increasing the number of classes, we proposed increasing the number of latent variables. We showed that because of the imposed constraints, the basic LC factor model with R latent variables has the same number of parameters as the LC cluster model with R+1 classes. This is an important result because it shows that in terms of parsimony, increasing the number of factors is equivalent to increasing the number of clusters.

The examples showed that in most cases the LC factor model provides a more parsimonious and easier to interpret description of the data. There is a simple explanation for this phenomenon. When applying a LC cluster model it is not known how many dimensions the solution will capture: A 3-cluster model may describe either one or two dimensions, while a 4-cluster model may describe either one, two, or three dimensions. When a 3-cluster model describes *one* dimension, it is very probable that a 1-factor model with 3 or more levels will describe the data almost as well (see the depression example). When a 3-cluster model describes *two* dimensions, it has the disadvantage that it can not capture all four basic combinations – (low, low), (high, low), (low, high) and (high, high) – of the two latent dimensions. Therefore, the 2- factor model will fit better than the 3-cluster model in these cases. In situations in which the 4-cluster model gives a 2-dimensional solution (as in the rheumatic arthritis data set where the 4 clusters represent the 4 possible latent combinations), it can be expected that a restricted 4-cluster model (the 2-factor model) will fit the data almost as well (and may be better in terms of BIC or p-value).

The above explanation yields strong arguments for using the two approaches in combination with one another, as we have been doing in the examples. There are two things that can happen when switching from the cluster to the factor model. First, the factor model may give a simpler description of the data than the cluster model. This occurs when the 3-cluster solution is one dimensional or when the 4-cluster solution is two dimensional, both of which are situations where the LC cluster model is overparametrized. Second, the factor model may give a better fit. We saw that this occurs when the three-cluster model is two-dimensional.

REFERENCES

Clogg, Clifford C. 1981. “New developments in latent structure analysis.” Pp. 215-246 *Factor analysis and measurement in sociological research*, edited by D.J. Jackson and E.F. Borgotta. Beverly Hills: Sage Publications.

Clogg, Clifford C. 1988. “Latent class models for measuring.” In *Latent trait and latent class models*, edited by R. Langeheine and J. Rost, New York, London: Plenum Press.

De Leeuw, Jan., and Peter G.M. Van der Heijden. 1991. “Reduced rank models for contingency tables.” *Biometrika* 78:229-232.

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

Gilula, Zri, and Shelby J. Haberman. 1986. “Canonical analysis of contingency tables by maximum likelihood.” *Journal of the American Statistical Association* 81:780-788.

Goodman, Leo A. 1974a. “Exploratory latent structure analysis using both identifiable and unidentifiable models.” *Biometrika* 61: 215-231.

Goodman, Leo 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.

Gower, John C., and David J. Hand. 1996. Biplots. London: Chapman & Hall.

Greenacre, Michael .J. 1993, *Correspondence analysis*. London: Academic Press.

Haberman, Shelby J. 1979. *Analysis of qualitative data, Vol 2, New developments*. New York: Academic Press.

Hagenaars, Jaques A. 1990. *Categorical longitudinal data – loglinear analysis of panel, trend and cohort data.* Newbury Park: Sage.

Hagenaars, Jaques A. 1993. *Loglinear models with latent variables*. Newbury Park: Sage.

Heinen, T. 1996. *Latent class and discrete latent trait models: similarities and differences.* Thousand Oaks: Sage Publications.

Hunt, Lyn and Murray Jorgensen. 1999. "Mixture model clustering using the MULTIMIX program." *Australian and New Zeeland Journal of Statistics* 41:153-172.

Lawrence C.J., W.J. Krzanowski. 1996. "Mixture separation for mixed-mode data. " *Statistics and Computing* 6:85-92.

Lazarsfeld, Paul F., and Neal W. Henry. 1968. *Latent structure analysis*. Boston: Houghton Mill.

Magidson, Jay. and Vermunt, Jeroen K. and 2001 (forthcoming). "Latent Class Factor and Cluster Models, Bi-plots and Related Graphical Displays", chapter in Sociological Methodology, Cambridge: Blackwell.

McCutcheon, Allan .L. 1987. *Latent class analysis*, Sage University Paper. Newbury Park: Sage Publications.

McLachlan, Geoffrey J., and Kaye E. Basford. 1988. *Mixture models: inference and application to clustering*. New York: Marcel Dekker.

Moustaki, Irini. 1996. "A latent trait and a latent class model for mixed observed variables. " *The British Journal of Mathematical and Statistical Psychology* 49:313-334.

Pearlin, Leonard I. and Joyce S. Johnson. 1977. "Marital status, life-strains, and depression." *American Sociological Review* 42:104-15.

Schaeffer, Nora C. 1988. "An application of item response theory to the measurement of depression", Pp. 271-308 in *Sociological Methodology 1988*, edited by C. Clogg. Washington DC: American Sociological Association.

Uebersax, J.S. 1993. Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, 88, 421-427

Van der Ark, L. Andries and Peter G.M. Van der Heijden. 1998. “Graphical display of latent budget and latent class analysis.” Pp. 489-509 in *Visualization of categorical data*, edited by J. Blasius and M. Greenacre. Boston: Academic Press.

Van der Ark, L. Andries, Peter G.M. Van der Heijden and D. Sikkel. 1999. “On the identifiability in the latent budget model.” *Journal of Classification* 16:117-137.

Van der Heijden, Peter G..M. Gilula, Z. and L. Andries Van der Ark. 1999 “On a Relation Between Joint Correspondence Analysis and Latent Class Analysis.” Pp. 147-186 in *Sociological Methodology 1999*, edited by M. Sobel and M. Becker.

Vermunt, Jeroen K. 1997. *LEM: A general program for the analysis of categorical data. User’s manual.* Tilburg University, The Netherlands.

Vermunt, Jeroen K. and Jay Magidson. 2000. *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.

Vermunt, Jeroen K. and Jay Magidson. 2001. “Latent Class Cluster Analysis”, Chapter 3 in *Applied Latent Class Analysis.* edited by J.A. Hagenaars and A.L. McCutcheon , Cambridge University Press.

Wolfe, John H. 1970. “Pattern clustering by multivariate cluster analysis. ” *Multivariate Behavioral Research* 5:329-350.