Nonloglinear Marginal Latent Class Models

Jacques A. Hagenaars
Wicher Bergsma
Marcel Croon

1. Introduction

The basic ideas underlying latent class analysis have been developed by Paul Lazarsfeld in the fifties of the previous century (Lazarsfeld 1950, 1959; Lazarsfeld and Henry 1968). The methods proposed by him and his coworkers for estimating the parameters of the latent class model were, although leading to BAN estimates, rather arbitrary and often resulted in very capricious or even impossible estimates. Despite the fact that in that period also the maximum likelihood equations for the latent class model were developed (see Lazarsfeld and Henry 1968), no practicable algorithms were available to obtain these maximum likelihood estimates on the desk calculators and low-speed computers then available. That situation improved spectacularly in the nineteen seventies thanks to the important, original contributions by Goodman and Haberman (Goodman 1974a,b; Haberman 1977, 1979, 1988). They were the first to develop feasible algorithms for obtaining maximum likelihood estimates for the basic latent class model and for a large number of important variants of the basic model. They truly converted Lazarsfeld's much promising latent structure model into a very flexible tool that could actually be used in practical research.
Building upon the work of Goodman and Haberman, many other people have since then made important contributions by enlarging the scope of the standard latent class model in some important statistical way and by showing how the latent class approach can be used to solve important substantive problems (for an applied methodological overview, see Hagenaars and McCutcheon 2002). Exemplary among those 'many other people' is certainly Dayton, who, often in collaboration with Macready, wrote numerous inventive and important applications of latent class analysis, several introductory articles and books on the topic and enlarged the latent class model significantly by showing how to incorporate continuous covariates into the model to explain the scores on the latent variable (e.g., Dayton 1991, 1998; Dayton and Macready 1983, 1988; Macready and Dayton 1992). In all this work, he always kept a keen eye on what substantive researchers needed. It is in this Daytonean spirit that we have tried to write this chapter: enlarging the basic latent class model in a manner that will be useful for substantive researchers.

First, in the next Section, a simple latent class model with two latent variables will be introduced. It will be shown how particular obvious and natural parameterizations and interpretations of the relationships in this model may result in a latent class model that can no longer be estimated by the standard methods developed by Goodman and Haberman implemented in most relevant statistical software. Alternative methods are required and are available.
In the last Section, a similar story will be told, but now using a somewhat more complicated latent class model, that is, a. Structural Equation Model (SEM) for categorical data emphasizing difficulties and possibilities somewhat different from the previous example.

2. Latent Class Models

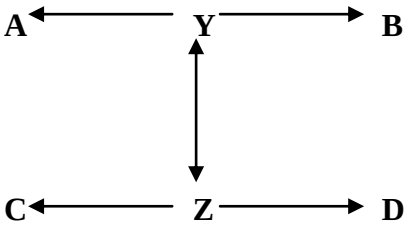
By way of example, Lazarsfeld's 1940 classical data set on Party and Candidate Preference in Erie County, Ohio will be used (Lazarsfeld 1972, p. 392). This data set is presented in Table 1. In this table, Party Preference is a dichotomous variable: 1.Democrats 2. Republicans, as is Candidate Preference: 1. Against Willkie (further indicated as Democrats) 2. For Willkie (further denoted as Republicans), where it must be remembered that Willkie was the (defeated) 1940 Republican Presidential Candidate running against Roosevelt. Hagenaars (1993) fitted several latent class models to the data in Table 1, among them the model that is presented in Figure 1.

Table 1 Party Preference (PP) and Presidential Candidate Preference (CP); Erie County Ohio, 1940; t_1 – August, t_2 – October

***A*. PP-t_1**	***B*.PP-t_2**	***C*. CP-t_1** ***D*. CP-t_2**	**1. Dem.** **1. Dem.**	**1. Dem.** **2. Rep.**	**2. Rep.** **1.Dem.**	**2. Rep.** **2. Rep.**
1.Dem	**2.Rep.**		68	2	11	12
1. Dem.	**2.Rep.**		1	1	0	1
2.Rep.	**1.Dem**		1	0	2	1
2. Rep.	**2.Rep.**		23	11	3	129

Source: Lazarsfeld 1972, p. 392

Figure 1



In the model in Figure 1, variable Y represents a dichotomous latent variable, indicating the (stable) underlying, 'true' Party Preference (1. Democrats 2. Republicans). Dichotomous latent variable Z is the (stable) latent Presidential Candidate Preference (1. Democrats 2. Republicans). Manifest variables A through D are the observed variables as denoted in Table 1. The model in Figure 1 can be written in terms of the following basic latent class (LCA) equation, where π_{yz}^{YZ} represents the joint probability of scoring (y,z) on YZ, $\pi_{a\,y}^{A|Y}$ the conditional probability of scoring A=a, given Y=y, and the other symbols have obvious analogous meanings.

$$\pi_{yzabcd}^{YZABCD} = \pi_{yz}^{YZ}\pi_{abcd\;yz}^{ABCD|YZ} = \pi_{yz}^{YZ}\pi_{a\,y}^{A|Y}\pi_{b\,y}^{B|Y}\pi_{c\,z}^{C|Z}\pi_{d\,z}^{D|Z} \qquad (1)$$

The first part of Eq. (1) ($\pi_{yzabcd}^{YZABCD} = \pi_{yz}^{YZ}\pi_{abcd\;yz}^{ABCD|YZ}$) is a tautology and by definition true, as follows from basic rules of probability calculus. However, the joint conditional probability $\pi_{abcd\;yz}^{ABCD|YZ}$ can be written in a more simple way as the product of the marginal conditional probabilities in the last part of Eq. (1) ($\pi_{yz}^{YZ}\pi_{a\,y}^{A|Y}\pi_{b\,y}^{B|Y}\pi_{c\,z}^{C|Z}\pi_{d\,z}^{D|Z}$), if the conditional independence restrictions implied by the model in Figure 1 are true. Central among these independence restrictions are the basic LCA assumption of conditional (local) independence among the indicators, given the scores on the latent variables.
Using the customary short hand notation to indicate hierarchical loglinear models, the model in Figure 1 can equivalently be represented as loglinear model {YZ,YA,YB,ZC,ZD, written out in full as

$$\ln \pi_{yzabcd}^{YZABCD} = \lambda + \lambda_y^Y + \lambda_z^Z + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_d^D + \lambda_{yz}^{YZ} + \lambda_{ya}^{YA} + \lambda_{yb}^{YB} + \lambda_{zc}^{ZC} + \lambda_{zd}^{ZD} \qquad (2)$$

This two latent variable latent class model fits the data in Table 1 well with (maximum likelihood-)$G^2 = 7.32$, df = 4 (p = .120, Pearson-$X^2 = 11.53$). The ML estimates of the parameters in Eq. (1) are given in Table 2.

Table 2 Estimates of Parameters in Eq. (1), applied to the data in Table 1

Y=y	**Z=z**	$\hat{\pi}^{YZ}_{yz}$	$\hat{\pi}^{A\|Y}_{1\,y}$	$\hat{\pi}^{A\|Y}_{2\,y}$	$\hat{\pi}^{B\|Y}_{1\,y}$	$\hat{\pi}^{B\|Y}_{2\,y}$	$\hat{\pi}^{C\|Z}_{1\,z}$	$\hat{\pi}^{C\|Z}_{2\,z}$	$\hat{\pi}^{D\|Z}_{1\,z}$	$\hat{\pi}^{D\|Z}_{2\,z}$
1	**1**	.315	.965	.035	.991	.009	.853	.147	.986	.014
1	**2**	.051	.965	.035	.991	.009	.081	.919	.000	1.000
2	**1**	.101	.013	.987	.004	.996	.853	.147	.986	.014
2	**2**	.534	.013	.987	.004	.996	.081	.919	.000	1.000

Latent class outcomes always contain a lot of detailed and interesting information. It is seen for instance in Table 2 that the true Republicans (Y=2 or Z=2) always answer just a bit more 'truthfully', i.e., have a somewhat larger probability of giving manifest Republican answers than the true Democrats who have a bit smaller chance of giving the 'correct' Democratic answer. Further the probability of giving the correct answer in agreement with the true position is always a bit smaller for the first wave in October than for the second wave in November. As Goodman has shown the statistical significance of these (small) differences can be investigated by testing the hypotheses that pertinent conditional response probabilities are equal to each other (Goodman 1974a,b).

Most attention will be paid here to the 'factor loadings', representing the associations between the latent variables and their indicators and thus expressing the 'reliability' of the measurements, assuming there is a one-to-one correspondence between the meanings of the categories of the latent variables and their indicators (Hagenaars 1990, 2002, 2005, 2010). Below, these 'factor loadings' or associations between the latent variables and their indicators will often be denoted in a general sense as 'reliabilities' without reference to the more restricted meaning of the term that is common for continuous variables, e.g., in classical error theory. The strength and direction of these associations or reliabilities can be expressed by means of the two-variable loglinear parameters in Eq (2), essentially by means of the the (log)odds ratios, that can be estimated on the basis of the conditional response probabilities in Table 2:

$$\hat{\lambda}_{11}^{YA} = 1.916,\ \hat{\lambda}_{11}^{YB} = 2.567,\ \hat{\lambda}_{11}^{ZC} = 1.046,\ \hat{\lambda}_{11}^{ZD} = 4.967$$

In general, the positive associations between the latent and the manifest variables and therefore the 'reliabilities' of the measurements are very high. According to these loglinear association coefficients, manifest variable D is the most reliable indicator, followed by B, A, and C. Note that the size of the effect of Z on D must just be regarded as 'very large'. Table ZD contains an almost empty cell making the odds ratio very sensitive to rounding errors. For example, using a few more digits after the decimal point in the calculations of the estimated conditional response probabilities and the loglinear parameter, the resulting estimate was $\hat{\lambda}_{11}^{ZD} = 18.898$. And to further illustrate the instability of the large effect: the above estimate $\hat{\lambda}_{11}^{ZD} = 4.967$ has an estimated standard error of 143.651.

The association between the underlying true Party Preference and the true Candidate Preference is found in table YZ with entries $\hat{\pi}_{yz}^{YZ}$ (Table 2). This association between the latent variables $\hat{\lambda}_{11}^{YZ} = .874$ corrects the corresponding observed relationships for misclassifications or measurement errors. (The observed relationships are for observed marginal table AC: $\hat{\lambda}_{11}^{AC} = .612$ and for table BD: $\hat{\lambda}_{11}^{BD} = .840$.) The estimated loglinear association $\hat{\lambda}_{11}^{YZ} = .874$ corresponds with an odds ratio of 33.036 and shows that the two latent variables are strongly positively related to each other. The odds of preferring the Democratic presidential candidate rather than the Republican one are 33 times higher for the Democratic Party supporters than for those who prefer the Republican Party.

However, expressing the directions and strength of the relationships among the variables in terms of the loglinear parameters (i.e., in terms of odds and odds ratios) is

in a way arbitrary. Other association measures as functions of the (conditional response) probabilities can be and have been used. For example, many researchers would prefer to describe the relationships between the latent variables and their indicators in terms of the differences ε between particular conditional response probabilities rather than in terms of ratios. For example, the effect of Y on A can also be estimated as follows, using the estimated conditional response probabilities in Table 2: $\hat{\varepsilon}_{11}^{A|Y} = \hat{\pi}_{11}^{A|Y} - \hat{\pi}_{12}^{A|Y} = .965 - .013 = .952$. In terms of ε, the effects of the latent variables on the indicators, i.e., the reliabilities of all indicators are:

$$\hat{\varepsilon}_{11}^{A|Y} = .952,\ \hat{\varepsilon}_{11}^{B|Y} = .987,\ \hat{\varepsilon}_{11}^{C|Z} = .772,\ \hat{\varepsilon}_{11}^{D|Z} = .986$$

Indicator C would now again be characterized as the most unreliable indicator, but the other indicators show more or less the same degree of reliability.
Also the relationship between the latent variables can be expressed in terms of ε using the estimated entries of table YZ (see Table 2) and arbitrarily treating Y as the independent variable. A rather strong relationship is found:
$\hat{\varepsilon}_{11}^{Z|Y} = .315/.366 - .101/.634 = .862 - .159 = .703$. The probability that someone prefers the Democratic presidential candidate is .70 higher for the Democratic Party than for the Republican Party supporters.
Although the above conclusions drawn from the odds ratios overlap to a large extent with the conclusions on the basis of the ε's, they are not exactly identical. The question then is what parameterization or association coefficient has to be preferred to express the relationships in the two latent variable latent class model: odds ratios, ε's or still other measures of association? Regardless of the specific model (but see below), a general textbook answer might be: choose that parameterization that suits your substantive theories best. However, most if not all social science theories lack this kind of precision, which is required to really guide the choice of the association coefficient. In general, the choice of a particular parameterization is much more a matter of taste and tradition and will be based on some general considerations of the properties of the coefficients. Many people find working with differences between conditional probabilities or percentages easier or more natural than applying odds (ratios); odds ratios have the nice property of being (variation) independent of the marginal distributions but are very sensitive to almost empty cells; epsilons, equivalent to unstandardized regression coefficients are much less sensitive to such an almost empty cell, but often cannot reach their theoretical maximum value given the marginal distributions, etc. etc.
But also, and most importantly in this chapter, the characteristics of a particular latent class model may restrict the range of association coefficients or parameterizations to choose from. The model in Figure 1 and Eq (1) is a graphical model (Whittaker 1990, Cox and Wermuth 1996), more precisely a DAG, a directed acyclical graph that is completely defined by the (conditional) independence relationships it implies among the variables. As long as the chosen parameterization of the latent class model has a one-to-one correspondence with these independence restrictions, it yields identical estimates and is admissible in this sense. If one chooses to define the basic latent class model in Eq (1) as a multiplicative (or loglinear) model (as in Eq (2)) and estimates the direction, strength, but also the absence of a direct relationship in terms of the presence or absence of particular (conditional) odds ratios, then the estimated probabilities for the entries of the complete table (here: table YZABCD) will be exactly the same as for the graphical model in Eq (1). This is true because statistical

independence implies that the pertinent log odds ratio equals zero (is 'absent'), but also a (n 'absent') log odds ratio of zero implies statistical independence between the pertinent variables. The same is true if one would write the basic latent class model in Eq (1) as an additive model and uses ε's to indicate the strength, direction but also the absence of a direct relationship among the variables, again because statistical independence implies $\varepsilon = 0$ and $\varepsilon = 0$ implies statistical independence.

However, as well known, all this does not apply to all association coefficients. For example, for larger than 2x2 tables (and nonnormal distributions), coefficients like Goodman and Kruskal's gamma (γ) or the product-moment correlation coefficient r are zero when there is statistical independence, but in general a value of zero of these coefficients does not imply statistical independence. Estimating a model such as the LCA model in Figure 1 under the restriction that, wherever an arrow is absent, a particular (conditional) γ or r is set to zero will therefore yield estimated probabilities for the complete table that will be different from the ones obtained when the model in Eq (1) is estimated.

Sometimes, researchers estimate a model such as the one in Eq (1) on the basis of the implied (conditional) independence relationships and then to use a coefficient such as (conditional) γ or r to describe the nature of the remaining relationships. Of course, technically it can be done, but we feel that it is at least strange and hard to defend from a theoretical, substantive point of view to define the absence of a relationship by means of a parameterization that is different from defining its strength and direction within the same model.

The explicit choice of an appropriate parameterization becomes more urgent and even necessary if (additional) restrictions are imposed on the LCA model that cannot be represented in the form of conditional independence relationships. For example, it is an obvious and natural research question to ask whether or not the reliabilities of the indicators in the above example are all the same in the population. But then it does matter for the test outcomes and the estimates how the reliabilities are expressed. In general, if the (log) odds ratios for two tables are the same, the ε's will be necessarily different and vice versa (confining ourselves from here on to these two coefficients). Therefore, estimating the probabilities for the complete table under the usual independence restrictions plus the extra restriction of equal reliabilities will yield different outcomes when the pertinent odds ratios (two-variable loglinear parameters) have been set equal to each other or the pertinent ε's.

Setting the reliabilities equal to each other in terms of equality restrictions on the odds ratios or loglinear parameters poses no special problems in the sense that such restrictions can easily be tested and the restricted reliabilities estimated using Haberman's and Goodman's procedures as implemented in widely used software such as LEM, MPLUS or Latent Gold (Muthen and Muthen 2006; Vermunt 1997b; Vermunt and Magidson 2005). (And as a side remark: because equalities of particular conditional response probabilities can be formulated in terms of restrictions on one- and two-variable loglinear parameters, also such 'reliability restrictions' pose no special problems (Goodman 1974a,b; Hagenaars 1990; Heinen 1996)).

However, for estimating latent class models with equal reliabilities in terms of ε's these standard estimation procedures cannot be used. Such a restriction of the reliabilities in terms of ε's brings the latent class model outside the exponential family because of which the standard (Goodman/Haberman) routines can no longer be used. However, an appropriate ML estimation procedure is provided by the marginal modeling approach. Becker and Yang were the first to extend marginal modeling to include latent variables and Bergsma advanced their procedure (Becker and Yang

1998; Bergsma 1997; Bergsma and Croon 2005; Bergsma and Rudas 2002a,b; Bergsma, Croon, Hagenaars 2009). Their algorithms are extensions of a very general method by Aitchinson and Silvey for finding MLE's for a very broad class of restrictions, further developed by Lang and Agresti (Aitchinson and Silvey 1958; Lang 1996; Lang and Agresti 1994; Lang, McDonald, Smith 1999). An extensive description of the marginal modeling approach with numerous real world applications including latent variabl emodels can be found in Bergsma, Croon, and Hagenaars (2009). Bergsma and Van der Ark wrote the program CMM (Categorical marginal Modeling), a flexible set of R (and Mathematica) routines for general marginal modeling, to be found at the webpage www.cmm.st. (Bergsma and Van der Ark 2009).

Bergsma, Croon, and Hagenaars (2009, Chapter 6) applied these marginal modeling procedures to the latent class model in Figure 1 for the data in Table 1, investigating several extra restrictions regarding the reliabilities of the indicators in terms of both odds ratios and ε's (and also in terms of Cohen's κ). The most restrictive hypothesis that all reliabilities in the two-latent variable model are the same has to be rejected both for the pertinent odds ratios ($G^2 = 25.16$, df = 7, p = .001) as for the ε's ($G^2 = 32.98$, df = 7, p = .000). The test result for the baseline two latent variable model without extra reliability restrictions discussed before was $G^2 = 7.32$, df = 4, p=.120. The 'all reliabilities equal' models can be conditionally tested against this baseline model, leading clearly to the same conclusions as the unconditional tests: they have to be rejected.

An interesting hypothesis that fits the data for the reliabilities in terms of odds ratios ($G^2 = 7.64$, df = 5, p=.177) but not in terms of ε's ($G^2 = 15.14$ df = 5, p = .005) is the restriction that in the two-latent variable model, the reliabilities increase from wave one to wave two, but with the same amount for party and candidate preference:

$$\lambda_{11}^{YA} - \lambda_{11}^{YB} = \lambda_{11}^{ZC} - \lambda_{11}^{ZD} \quad or$$
$$\varepsilon_{11}^{YA} - \varepsilon_{11}^{YB} = \varepsilon_{11}^{ZC} - \varepsilon_{11}^{ZD}$$

In terms of ε as reliability measure, a model in which the reliabilities of party preference were supposed not to change:

$$\varepsilon_{1\,1}^{A|Y} = \varepsilon_{1\,1}^{B|Y}$$

but change was allowed in the reliability of candidate preference fitted the data nicely: $G^2 = 8.45$, df = 5, p = .133. The reliabilities were estimated as

$$\hat{\varepsilon}_{1\,1}^{A|Y} = \hat{\varepsilon}_{1\,1}^{B|Y} = .969 \quad (s.e..012)$$
$$\hat{\varepsilon}_{1\,1}^{C|Z} = .774 \quad (s.e..042)$$
$$\hat{\varepsilon}_{1\,1}^{D|Z} = .981 \quad (s.e..023).$$

Different conclusions can and sometimes will/must be reached when different parameterizations are applied. For many models and for many research questions, researchers may have good reasons to express their hypotheses about the reliabilities of the indicators not in terms of odds and odds ratios but to use other effect measures such as percentage differences. The application of marginal modeling procedures makes it possible, given further development of the software to routinely apply such

nonloglinear parametrizations in latent class models. Other instances where this might be useful will be discussed in the next section in the context of Structural Equation Models for categorical data.

3. Structural Equation Models

The principles of SEMs - Structural Equation Models for categorical data have been outlined a long time ago (Goodman 1973a,b). Since then, Goodman's models have been extended to include categorical latent variables (Hagenaars 1990, 1993, 1998, 2002; Hagenaars, Heinen, Hamers 1980; Vermunt 1997a), they have been integrated into the general framework of graphical modeling (Pearl 2000; Cox and Wermuth 1996) and user friendly software has been developed to routinely estimate these models (LEM, Latent Gold); Vermunt 1997b; Vermunt and Magidson 2005). In this section, further useful extensions will be discussed using marginal modeling procedures (see also Croon, Bergsma, Hagenaars 2000; Bergsma, Croon, Hagenaars 2009, Chapters 5,6). The data that will be used are presented in Table 3. They are very much like the data set in the previous section, in the sense that the observed variables represent Party and Candidate Preference in two waves. But these data are from the Dutch population (and much later) and are trichotomies: 1.Christian Democrats 2. Left wing 3. Other (mainly Right wing). For the interpretation of the results, it must be remembered that in the Dutch political system, 'candidate' refers the candidate for Prime-Minister, who is not elected by the voters, unlike the president in the USA.

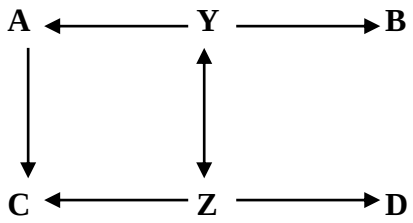
Table 3 Party and Candidate Preference (Source Hagenaars 1990)

		C	*1*	*1*	*1*	*2*	*2*	*2*	*3*	*3*	*3*	
		D	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*	*Total*
A	B											
1	*1*		84	9	23	6	13	7	24	8	68	242
1	*2*		0	1	0	0	8	1	2	2	3	17
1	*3*		3	1	2	0	2	3	2	3	9	25
2	*1*		1	1	0	1	2	2	1	0	1	9
2	*2*		2	4	0	1	293	6	1	22	21	350
2	*3*		1	0	0	1	8	7	0	0	9	26
3	*1*		6	1	1	4	5	0	9	1	16	43
3	*2*		0	1	1	0	31	0	2	9	7	51
3	*3*		14	1	15	3	48	23	12	21	200	337
Total			111	19	42	16	410	49	53	66	334	1100

A - Party Preference t_1
B - Party preference t_2
C - Candidate Preference t_1
D - Candidate Preference t_2

1. Christian Democrat
2. Left Wing
3. Other

Figure 2



The same model as before, that is, the model in Figure 1 equivalently represented in the shorthand loglinear notation as model {YZ,YA,YB,ZC,ZD} can be fitted to the data in Table 3, treating the latent variables also as trichotomous variables. However, the model does not fit the data (G^2 = 84,74, df = 48, p = .00 (Pearson X^2 = 94.33)) (although better than some competing latent class models; Hagenaars 1990, 2010; Bergsma, Croon, Hagenaars 2009). A possible reason might be that people did not have a clear idea in the first wave about the possible candidates for Prime Minister and were influenced in their answer to C-Candidate Preference at time 1 not only by their true Candidate Preference but also by their previous answer within the same interview regarding their Party Preference (Variable A). This is a kind of test-retest or consistency effect which violates the local independence assumption of the basic latent class model and results in a local dependence model (Hagenaars 1988, 1990, 2010; Bassi, Croon, Hagenaars, Vermunt 2000). A possible model along these lines is is depicted in Figure 2.

The model in Figure 2 model can no longer be represented by one particular loglinear model, e.g., model (YZ, YA,YB,ZC,ZD,AC}. In a particular loglinear model (as in a multiple regression model), all effects are partial effects controlling for the other variables in the model. For example, the direct relationship between Y and Z (λ_{yz}^{YZ} in model (YZ,YA,YB,ZC,ZD,AC}) is the relationship between the latent variables, controlling for C (and other variables). But according to the 'causal' diagram in Figure 2, C is a consequence of Z and 'causally later' than Y and Z and can therefore not influence Y or Z or their relationship. Given the causal order, the relation between Y and Z must be investigated in marginal table YZ, collapsing over the other variables. (In a way this is also true regarding the model in Figure 1, but there the collapsibility theorem guarantees that the results for Y-Z are the same whether or not controlling for the other variables; this is no longer true here because of the direct effect of A on C. In graphical terms, it amounts to whether or not a directed graph implies the same independence restrictions as the corresponding undirected graph (Bishop, Fienberg, Holland 1975; Whittaker 1990; Lauritzen 1996). As Goodman showed and as follows from the rules of directed graphical modeling, if the model in Figure 2 is true, including the (causal) order among the variables and including the implied (conditional) independence restrictions, the joint probability π_{yzabcd}^{YZABCD} can be written as follows

$$\pi_{yzabcd}^{YZABCD} = \pi_{yz}^{YZ}\pi_{ay}^{A|Y}\pi_{by}^{B|Y}\pi_{caz}^{C|AZ}\pi_{dz}^{D|Z} \qquad (3)$$

In order to obtain the estimates for the right hand side elements of Eq (3) and the appropriate effect parameters, a loglinear submodel has to be defined for each of these elements. In agreement with the model in Figure 2, a saturated loglinear submodel will be applied to marginal table YZ with entries π_{yz}^{YZ}, a saturated logit submodel to marginal table YA (with A as the dependent variable), a saturated logit submodel to marginal table YB, and a saturated logit submodel to marginal table ZD. To marginal table ZAC, a nonsaturated logit submodel must be applied, because it is assumed here that A and Z do not interact regarding their influence on C. In terms of loglinear models, submodel {AZ,AC,ZC} must be fitted to marginal table ZAC. These submodels provide the correct parameter estimates for the relationships among the variables, correct in the sense of taking the causal order of the variables into account. Compared to ordinary SEMs for continuous data where the SEM consists of a set of

multiple regression equations, here a 'corresponding' set of loglinear or logit equations is used.
The submodels also provide the estimates for the right hand side elements in Eq (3) and from them the left hand side probability can be calculated. The estimates $\hat{\pi}^{YZABCD}_{yzabcd}$ are the estimated entries for the complete table, given that all (conditional) independence restrictions implied by the model in Figure 2 and all additional assumptions implied by the loglinear submodels for the right hand side elements are true. A test that the whole causal model with all its implications and restrictions is true can be obtained by summing $\hat{\pi}^{YZABCD}_{yzabcd}$ over the latent variables and comparing the resulting probabilities in the usual way with the observed frequencies in table ABCD (Table 3). The model in Figure 2 turns out to fit the data in Table 3 very well: : G^2 = 45.97, df = 44, p = .39 (Pearson X^2 = 44.04).
So far, this categorical SEM approach is fairly standard by now and can be routinely applied by a program such as LEM. However, again, there might be a number of additional interesting research questions that cannot be answered by this standard approach. The first one may have to do (again) with the particular parameterization chosen. Above it was assumed that A and Z did not interact regarding their influence on C. Therefore loglinear model {AZ,AC,ZC} was fitted to marginal table ZAC. This model parameterizes the no-three-variable-interactions in loglinear terms. It is assumed that corresponding conditional odds ratios for the effect of A an C are the same in all three categories of Z (or equivalently the odds ratios for Z-C are the same for all three categories of A). But, as above, a researcher might be more interested in an additive parameterization rather than a multiplicative (or loglinear) one and may want to use ε as effect measure. Imposing the restrictions of no-three-variable-interaction in terms of ε's will yield results that are different from using odds ratios (and cannot be carried out by the standard Goodman procedure). Although many details of defining a SEM such as the one in Figure 2 as an additive model (additive in the frequencies rather than the log frequencies) have still to be worked out, the marginal modeling procedures referred to above provide an excellent starting point for estimating such additive SEMs for categorical data.
Another research question that must be answered by means of marginal modeling concerns a marginal homogeneity hypothesis at the latent level (see also Hagenaars 1986). For many theoretical and practical reasons it might be interesting to know whether or not the true party preferences of the respondents differ from their true candidate preferences. The estimated marginals of table YZ for the model in Figure 2 are as follows, with subscript i = 1. Christian Democrats i = 2. Left Wing i = 3. Other:

Party Preference: $\hat{\pi}^{YZ}_{i+}$:*1.*.275 *2.*.374. *3.*.351
Candidate Preference: $\hat{\pi}^{YZ}_{+i}$:*1.*.193 *2.*.428. *3.*.379

Especially the Christian Democratic party is more popular than the (newly designated) Christian Democratic Candidate and just the opposite is true for Left wing. To investigate whether or not these differences are significant, the model in Figure 2 will be estimated but now under the extra marginal homogeneity restriction $\pi^{YZ}_{i+} = \pi^{YZ}_{+i}$; this again brings the model outside the exponential family and marginal modeling procedures must be used. The test outcomes for the whole (but MH restricted) model in Figure 2 are on the borderline of significance: G^2 = 62.33, df=46 (p=. 011, X^2 =57.10). However, the more powerful conditional test, testing the model

with against the model without the extra restriction provides a clear result: G^2 = 62.33-45.97 = 16.36, df = 48-46 = 2, p = .000. The hypothesis that the distributions of the true Party Preference and the true Candidate Preference are homogenous in the population has to be rejected and our best guess is that the differences are as described above.

Many more examples might be given to show the usefulness of enriching latent class modeling using marginal modeling procedures (and are given by Bergsma, Croon, Hagenaars 2009). Most importantly, this is not an enrichment just for the sake of enlarging the scope statistical modeling but it does provide researchers with even more powerful tools to answer important and often occurring research questions. C. Mitchell Dayton might approve!

References

Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, *29*, 813-828.

Bassi, F., Hagenaars, J. A., Croon, M. A., & Vermunt, J. K. (2000). Estimating true changes when categorical panel data are affected by uncorrelated and correlated errors. *Sociological Methods and Research*, *29*, 230-268.

Becker, M. P., & Yang, I. (1998). Latent class marginal models for crossclassifications of counts. In A. E. Raftery (Ed.), *Sociological methodology* (Vol. 28, p. 293-326). Oxford: Blackwell.

Bergsma, W. P. (1997). *Marginal models for categorical data.* Tilburg: Tilburg University Press.

Bergsma, W. P., & Van der Ark, L. A. (2009). CMM: Categorical marginal models. R package version 0.1.

Bergsma, W. P., & Croon, M. A. (2005). Analyzing categorical data by marginal models. In L. A. Van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (p. 83-101). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Bergsma, W.P., Croon M., and Hagenaars, J.A. (2009). *Marginal models for dependent, clustered, and longitudinal categorical data*. Berlin: Springer

Bergsma, W. P., & Rudas, T. (2002a). Marginal models for categorical data. *Annals of Statistics, 30,* 140-159.

Bergsma,W. P.,&Rudas, T. (2002b). Modeling conditional and marginal association in contingency tables. *Annales de la Facult´e des Sciences de Toulouse, XI*,

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975) *Discrete multivariate analysis; Theory and Practice.* Cambridge, Mass: MIT Press

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation.* London: Chapman & Hall.

Croon, M. A., Bergsma, W. P., & Hagenaars, J. A. (2000). Analyzing change in categorical variables by generalized log-linear models. *Sociological Methods and Research, 29*, 195-229.

Dayton, C.M. (1991) Educational applications of latent class analysis. *Measurement and evaluation in counseling and development, 24*, 131-141

Dayton, C. M. (1998) *Latent class scaling analysis*. Thousand Oaks: Sage

Dayton, C.M. and Macready, G.B. (1983) Latent structure analysis of repeated classifications with dichotomous data. *British journal of mathematical and statistical psychology, 36,* 189-210.

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

Goodman, L. A. (1973a). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika, 60,* 179-192.

Goodman, L.A. (1973b) Causal analysis of data from panel studies and other kinds of surveys. *The American Journal of Sociology,* Vol. 78, pp. 1135-1191

Goodman, L. A. (1974a). 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.

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

Haberman, S. J. (1977). Product models for frequency tables involving indirect observations. *Annals of Statistics, 5,* 1124-1147.

Haberman, S. J. (1979). *Analysis of qualitative data. vol.2, new developments*. New York: Academic Press

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for loglinear models for frequency tables derived by indirect observation. In C. C. Clogg (Ed.), *Sociological methodology: Vol. 18* (p. 193-211). Washington, D.C.: American Sociological Association.

Hagenaars, J. A. (1986). Symmetry, quasi-symmetry, and marginal homogeneity on the latent level. *Social Science Research, 15,* 241-255.

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: Log-linear, panel, trend, and cohort analysis.* Newbury Park: Sage.

Hagenaars, J. A. (1993). *Loglinear models with latent variables.* Newbury Park: Sage.

Hagenaars, J. A. (1998). Categorical causal modeling: latent class analysis and directed loglinear models with latent variables. *Sociological Methods and Research, 26,* 436-486.

Hagenaars, J. A. (2002). Directed loglinear modeling with latent variables: Causal

models for categorical data with nonsystematic and systematic measurement errors. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (p. 234-286). Cambridge: Cambridge University Press.

Hagenaars, J. A. (2005). Misclassification phenomena in categorical data analysis: Regression toward the mean and tendency toward the mode. In L. A. V. der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (p. 15-39). Mahwah, NJ: Lawrence Erlbaum.

Hagenaars J.A (2010, forthcoming) Loglinear latent variable models for longitudinal categorical data. In Van Montfort K., Oud, H, and Satorra A. (Eds.) *Longitudinal Research with latent variables*. Springer

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis.* Cambridge: Cambridge University Press.

Hagenaars J.A., Heinen, A.G.J. and Hamers P.A.M. (1980) Causale modellen met diskrete latente variabelen: Een variant op de LISREL-benadering. *Methoden en Data Nieuwsbrief*, vol 5, pp 38-54. VVS-Vereniging voor Statistiek; SWS-sociaal-wetenschappelijke Sectie.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Thousand Oaks CA: Sage.

Lang, J. B. (1996). Maximum likelihood methods for a generalized class of loglinear models. *Annals of Statistics, 24*, 726-752.

Lang, J. B., & Agresti, A. (1994). Simultaneously modelling the joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association, 89*, 625-632.

Lang, J. B., McDonald, J. W., & Smith, P. W. F. (1999). Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *Journal of the American Statistical Association, 94*, 1161-1171.

Lauritzen, S. L. (1996). *Graphical models.* Oxford: Clarendon Press.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. Stouffer (Ed.), *Measurement and prediction* (p. 413-472). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: a study of a science. conceptual and systematic. Vol. 3: Formulations of the person and the social context* (p. 476-543). New York: McGrawHill.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lazarsfeld P.F. (1972) The problem of measuring turnover. In P.F. Lazarsfeld, a.K. Pasanella and M. Rosenberg (Eds), *Continuities in the language of social research.* Pp. 388-398. New York: Free Press.

Macready, G.B. and Dayton, C.M. (1992) The application of latent class models in adaptive testing. *Psychometrika, 57*, 71-88.

Muthén, L. K., & Muthén, B. O. (2006). *Mplus: Statistical analysis with latent variables. (user's guide fourth edition)*. Los Angeles, CA: Muthén and Muthén.

Vermunt, J. K. (1997a). *Log-linear models for event histories.* Thousand Oaks, CA: Sage.

Vermunt, J. K. (1997b). *LEM: A general program for the analysis of categorical data: Users manual* (Tech. Rep.). Tilburg, NL: Tilburg University

Vermunt, J.K. and Magidson, J. (2005) Latent GOLD 4.0 User's Guide. Belmont, Massachusetts: Statistical Innovations Inc.

Whittaker, J. W. (1990). *Graphical models in applied multivariate statistics.* New York: Wiley.