Wicher P. Bergsma, Marcel A. Croon and Jacques A. Hagenaars

Advancements in marginal modeling for categorical data

Article (Accepted version)
(Refereed)

Original citation:
Bergsma, Wicher P., Croon, M. A. and Hagenaars, J. A. (2013) *Advancements in marginal modeling for categorical data.* Sociological Methodology, 43 (1). pp. 1-41. ISSN 0081-1750 DOI: 10.1177/0081175013488999

Advancements in Marginal Modeling for Categorical Data

Wicher Bergsma, Marcel Croon, Jacques Hagenaars

Abstract:

Very often the data collected by social scientists involve dependent observations, without, however, the investigators having any substantive interest in the nature of the dependencies. Although these dependencies are not important for the answers to the research questions concerned, they must still be taken into account in the analysis. Standard statistical estimation and testing procedures assume independent and identically distributed observations, and need to be modified for observations that are clustered in some way. Marginal models provide the tools to deal with these dependencies without having to make restrictive assumptions about their nature. In this paper, recent developments in the (maximum likelihood) estimation and testing of marginal models for categorical data will be explained, including marginal models with latent variables. The differences and commonalities with other ways of dealing with these nuisance dependencies will be discussed, especially with GEE and also briefly with (hierarchical) random coefficient models. The usefulness of marginal modeling will be illuminated by showing several common types of research questions and designs for which marginal models may provide the answers, along with two extensive real world examples. Finally, a brief evaluation will be given, shortcomings and strong points, computer programs and future work to be done.

1. INTRODUCTION

In social science research many interesting substantive theories and hypotheses are investigated by comparing different marginal distributions defined for an appropriate selection of variables rather than by looking at the properties of the total joint distribution for all variables involved in the data collection procedure. Studying agreement or differences among various marginal distributions is almost always based on tables that are not obtained from independent samples of respondents, but are derived from the same overall sample. As a consequence these tables may show varying degrees of dependency which has to be taken into account in the statistical analysis. In their book Bergsma, Croon, and Hagenaars (2009) described a maximum likelihood (ML) approach for testing hypotheses about marginal distributions, and estimating the relevant parameters in the corresponding models. In their approach the dependencies among the data are directly incorporated in the likelihood function itself, making any ad hoc specification of the potential dependencies in the data unnecessary. By applying these methods to data coming from a variety of social surveys, they showed how ubiquitous marginal models really are in substantive research in sociology. As we will frequently refer to Bergsma et al. (2009), we abbreviate this reference by BCH.

The main purpose of this paper is to further propagate this new methodology among an audience of social science researchers. In this first section the need for marginal modeling in social science research is demonstrated by a simple example, and the intuitive ideas behind estimation and testing are given. The second section is devoted to a more formal exposition of the maximum likelihood procedures described in BCH. It is outlined how missing data can be dealt with, which was not done in BCH. In the third section, other approaches are described, with particular attention given to the generalized estimating equations (GEE) and GSK (after Grizzle, Starmer and Koch, 1969) approaches. Similarities and dissimilarities with maximum likelihood estimation, as well as relative advantages and disadvantages, are discussed. It is highlighted how questions that can best be answered using marginal models differ from those that can best be answered using random coefficient models. In the fourth section the ML approach is applied to two concrete data sets, the first of which was not analyzed before using marginal modeling. In the first example marginal modeling is applied to data collected in a rotating panel design. The analysis here is carried out on tables which are partly dependent. This example shows that marginal modeling methods can easily be extended to data from complex sampling designs. In the second example, taken from Bergsma et al., a classical data set (Lazarsfeld, 1972) is analyzed by means of a latent class model in which both loglinear and non-loglinear constraints are imposed on the cell probabilities. These two non-trivial examples should make clear that the ML approach in marginal modeling is not restricted to relatively simple research questions, but remains applicable in much more complex circumstances, irrespective of whether these concern the sampling design or the structure of the statistical model.

1.1 General Characteristics

The oldest and best known marginal model is probably the Marginal Homogeneity (MH) model for square tables. Many of the characteristic features and uses of marginal modeling can be captured by means of the MH model and its straightforward extensions (Caussinus 1966; Grizzle, et al 1969; Bishop et al. 1975; Haberman 1979; Duncan 1979, 1981; Haber 1985; Hagenaars 1986, 1990).

The data in Table 1 are from a panel study, part of the US National Election Study, in which the same respondents are interviewed several times. Table 1 is a turnover table that represents the individual gross changes in Political Orientation (measured on a seven-point scale) in the US.

Table 1. Political Orientation (US national election studies)

	B.t₂ - 1994							
A.t₁ -1992	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Total*
1	3	4	1	2	0	1	0	11
2	2	23	15	6	0	2	0	48
3	1	8	23	9	9	1	0	51
4	0	6	17	56	19	13	2	113
5	0	1	1	18	40	29	3	92
6	0	1	1	4	13	51	7	77
7	0	0	0	0	2	11	3	16
Total	6	43	58	95	83	107	16	408

Source U.S. National Election Studies; see also Bergsma et al (2009)

1. extremely liberal 2. *liberal* 3. *slightly liberal* 4. *moderate*
5. *slightly conservative* 6. *conservative* 7. *extremely conservative*

The observed frequency entries $f_{i\,j}^{AB}$ in Table 1 can be used to estimate the joint probabilities $\pi_{i\,j}^{AB}$ in the population, or the conditional probabilities $\pi_{j\,i}^{B|A}$. In this way, the amount and nature of the individual changes can be investigated by looking at how an individual's position at Time 2 depends on the scores at Time 1. However, very often the research questions to be answered in these studies do not concern the individual gross changes but rather the overall net changes. Researchers will then use a table such as Table 1 to investigate the net changes in Political Orientation, comparing the marginal frequencies $f_i^A(= f_{i\,+}^{AB})$ and $f_j^B\left(= f_{+\,j}^{AB}\right)$. Typical research questions in this kind of study are: have people become more liberal or more conservative from Time 1 to Time 2? Or, has the population become less or more diverse regarding its Political Orientation from Time 1 to Time 2? For answering these questions, the patterns and dependencies within the turnover are irrelevant; only the differences between the marginal distributions provide the substantively relevant information.

In Table 2, the two marginal distributions from Table 1 are put together with a third wave added from the same US Election Study.

Table 2. Marginal Distributions of Political Orientation (US national election studies). Source: see Table 1.

	T-Time				
P-Political Orientation	*A.t_1-1992*	*B.t_2-1994*	*C.t_3-1996*	*MH*	*Ind.*
1 extr.lib	11	6	6	9.17	7.67
2	48	43	36	42.52	42.33
3	51	58	69	58.78	59.33
4 mod.	113	95	98	103.6	102.00
5	92	83	86	86.90	87.00
6	77	107	98	91.83	94.00
7 extr.cons.	16	16	15	15.21	15.67
Total	408	408	408	408	408

Marginal Homogeneity (model [T,P]): G^2 = 27.66, df = 12, p = .006 (X^2 = 26.11)
(Naïve) Independence (model [T,P]): G^2 = 14.04, df = 12, p = .298 (X^2 = 14.06)

*Note. Co*lumn *MH* contains maximum likelihood estimates of expected frequencies under marginal homogeneity hypothesis. Column *Ind.* contains expected frequencies under hypothesis of independent samples, i.e., the average frequency for the three time points.

The hypothesis that there is no (net) change in Political Orientation is equivalent to the independence hypothesis for the data in Table 2. Obviously, this hypothesis also implies that the marginals in Table 1 are identical to each other so that there is marginal homogeneity in Table 1. In terms of the standard short hand notation for denoting hierarchical loglinear models, the independence model [*T,P*] should be valid for the data in Table 2 with *T* representing Time and *P* Political Orientation. Its loglinear representation is

$$\ln \pi_{tp}^{TP} = \lambda + \lambda_t^T + \lambda_p^P.$$

Taking the dependencies in the data into account, the test of marginal homogeneity yields G^2 = 27.66, with df = 12, p = .006 (X^2 = 26.11), i.e., there is strong evidence that the marginal distributions change over time. However, a different result would be obtained if the dependencies are not taken into account. If Table 2 had been obtained by means of repeated cross-sections with three independent samples, it would have been a standard table *TP* (Time x Political Orientation) containing *iid* (independent and identically distributed) observations and the independence model could be tested by means of the standard chi-square procedures, yielding G^2 = 14.04, with df = 12, p = .298 (X^2 = 14.06). The conclusion would be that there is no reason to reject the hypothesis that the distribution of Political Orientation is the same for the three time points: there is no net change. This is true according to the maximum likelihood chi-square G^2 as well as Pearson-chi-square X^2. Given that the three sample sizes are equal (N = 408), the maximum likelihood estimates of the distribution under the independence assumption would have been the mean of the three distributions and would look like the entries in the last Column of Table 2.

However, Table 2 does not come from a trend (repeated cross-sectional) design but from a truly longitudinal (panel) study in which the same respondents are interviewed three times. The same respondents appear in all three distributions of Political Orientation. In other words, the data in Table 2 are not *iid*, which is a basic assumption underlying the standard chi-square test used above. The observations are clustered and dependent upon each other, as illustrated by the association patterns in Table 1. In general, such dependencies will seriously affect the standard errors of the estimates $\hat{\pi}$ of the cell probabilities, and accordingly the size of the test statistics. This is clearly seen from the row "Marginal Homogeneity" in Table 2. This row contains the chi-square statistics for testing the hypothesis that the three distributions are equal, but now taking the dependencies among the observations into account without imposing any restrictions on the dependence structure by using the marginal modeling maximum likelihood (ML) approach, which is more formally discussed in Section 2. The number of degrees of freedom does not change, but the values of the chi-square statistics are now much higher. Consequently, the hypothesis of equal distributions in the population must be rejected and the presence of net change should be accepted.

In a way, such a result is not unexpected (Hagenaars 1990, p. 206). When comparing the *t*-test for the difference between two means for independent and matched samples, the standard errors are smaller when the correlation among the observations is positive and the ensuing test statistics are higher (although in terms of probability levels, this is partially upset by the increase in degrees of freedom for the *t*-test in the independent case). On the other hand, if the covariance between the two sets of observations is negative, the t-values are expected to be lower in the matched case, and if the covariance is zero, the matched and the independent case produce equivalent results. A similar kind of reasoning applies here. However, --- and that is the reason to mention it explicitly --- one cannot simply say that ignoring the dependencies in the data always leads to lower values of the chi-square statistics, not even when the dependence structure is (seemingly) positive. Obviously, if the association pattern (as in Table 1) happens to agree with statistical independence, the appropriate test taking the dependencies among the observations into account and the naïve test ignoring the dependencies produce the same results. However, in more complicated, but real world applications, BCH provide many examples in which the naïve test for independence yields a higher or similar G^2–value than the correct test on marginal homogeneity despite the fact that the items all show positive associations for the two-way tables.

The Column MH in Table 2 also presents the maximum likelihood estimates for the distribution of Political Orientation under the assumption of marginal homogeneity while taking the dependencies into account. By comparing the last two columns in Table 2, it is immediately seen that the naïve estimates (Column Ind.) and the appropriate estimates (Column MH) are different. This is generally true when estimating restricted marginal models for categorical data. There are some known exceptions, e.g., when the dependencies among the observations show a symmetrical association pattern, as in the loglinear quasi-symmetry, uniform association, or independence models. But in general, the naïve estimates and the appropriate ML estimates for the distribution of categorical variables will be different. Although ignoring the dependence structure as in the naïve estimates in Column Ind. generally still provides consistent estimators (see the discussion on GEE below in the next subsection), these naïve estimators have higher (asymptotic) standard errors than the maximum likelihood estimators if marginal homogeneity holds, except in special cases such as independence when the standard errors are the same.

1.2 The Basic Approach

Before turning to the formal exposition of marginal modeling in Section 2, it might be helpful at least for some readers to first get a very rough, intuitive idea of the most basic elements of the estimation and testing principles involved.

The dependencies among the three distributions to be compared in Table 2 occur because the same individuals are involved in all three distributions. The observations over time are nested within individuals and the individual can be regarded as the clustering unit. A first requirement of the marginal estimation and testing procedure is that the clustering units (i.e., the individuals) are a random sample from the intended population. Confining the discussion to the first two time points in Table 2, Table 1 shows the dependencies among the observations for the (marginal) distributions of 1992 and 1994. If the individuals are indeed a random sample of the population and when no further restrictions are imposed on the entries of Table 1, the saturated model applies and the observed proportions $p_{i\,j}^{AB}(= f_{i\,j}^{AB}/\mathrm{N})$ can be used as the maximum likelihood estimates $\hat{\pi}_{i\,j}^{AB}$ for the corresponding probability $\pi_{i\,j}^{AB}$ in the population. A particular cell estimate $\hat{\pi}_{i\,j}^{AB}$ follows a multinomial distribution with estimated variance

$$\hat{\pi}_{i\,j}^{AB}\left(1-\hat{\pi}_{i\,j}^{AB}\right)/N.$$

The estimated covariance between two estimated cell probabilities in Table 1, say $\hat{\pi}_{i\,j}^{AB}$ and $\hat{\pi}_{i'j'}^{AB}$, equals

$$-\hat{\pi}_{i\,j}^{AB}\hat{\pi}_{i'j'}^{AB}/N.$$

Once the estimated (co)variances of the entries in the full joint table are known, it is rather straightforward to obtain the estimated (co)variances of (weighted) sums of cells and the chi square test statistics for contrasts between such sums.

If restrictions are imposed on the marginal probabilities, e.g., marginal homogeneity, the appropriate maximum likelihood estimates $\hat{\pi}_{i\,j}^{AB}$ must be obtained under this restriction and these are then used in the way indicated above to get the estimated (co)variances of the estimated rather than the observed proportions.

The restrictions may pertain to the marginal tables, but also to the dependencies in the joint table (Lang and Agresti 1994; Croon et al. 2000; Vermunt et al. 2001). In this way, one can estimate and test, for example, a model for Table 1 in which simultaneously marginal homogeneity is assumed for the marginals and a linear by linear (uniform) association for the turnover table itself. For previous work on this, see Bartolucci and Forcina (2002), who considered marginal models combined with RC models for the joint distribution. (Note that RC models for the joint distribution can be fitted using the same methodology as outlined in this paper, although some extra work is required because the likelihood needs to be reparameterized in terms of the RC model parameters. Unlike the linear by linear association model, the RC model is not in the natural exponential family, making the computations required for this reparameterization a bit more involved, see Bartolucci and Forcina for details.)

Interest need not be confined to comparing entire marginal distributions, but may be extended to functions of cell probabilities defined on such marginals. For example, in Table 1, one might be interested in comparing the sum of all frequencies above the main diagonal, (i.e. all cells indicating a tendency to be more conservative at Time 2 than at Time 1) with the sum of all cells below the main diagonal (indicating the tendency towards more liberalism). Or, one might investigate whether the significant net change in marginal distributions in Table 1 has to do with a net change in the intensity of the orientation (extreme to moderate) or with the shift in direction (liberal-conservative), each time summing the appropriate but different sets of cells. (The answer is that both tendencies are involved, see BCH, p. 106.) Alternatively, one might investigate whether the mean Political Orientation at Time 1 is different from the corresponding mean at time 2, or whether the variance or dispersion at Time 1 is different from the variation at Time 2.

Such functions of marginal probabilities are the more interesting if the marginal tables concern two or more variables. For example, assume that in a panel study next to the repeated measurements of Political Orientation (P), also Religiosity (R) is measured several times, along with Gender (G). This gives rise to a table $GR_1P_1R_2P_2$ (for two points in time). Research questions that require marginal modeling procedures would be whether or not the association between Gender and Political Orientation is the same at Time 1 as at Time 2; whether or not the association between Religiosity and Political Orientation has stayed the same for the two points in time; whether the latter holds true for both men and women, etc. Such questions can be answered by comparing the relevant two- and three dimensional marginal tables as a whole or in terms of suitable association coefficients (odds ratios, product moment correlations etc.).

It can become quite complicated, certainly when additionally categorical latent variables are involved, how to arrive from the original cell frequencies in the joint (full) table with their estimated (co)variances to these complex functions in the marginal tables, along with their (co)variances. In Section 2, the necessary matrix operations are presented, along with a nice tool: the generalized exp-log notation. This notation has been developed originally by Grizzle, et al., and further generalized by Bergsma (Grizzle et al 1969, Kritzer 1977, Bergsma 1997, BCH). With these tools, a large number of interesting types of research questions can be answered, some of which are presented in the next subsection.

1.3 Types of Research Questions Requiring Marginal Modeling

A very important area of application of marginal modeling is strictly longitudinal research with repeated measurements on the same respondents. Although it is always said – and we believe it to be true – that the great strength of longitudinal research is the study of individual gross changes, very often longitudinal data are simply used for investigating net changes or changes in marginal tables. To provide a few examples: panel data are used to study how the one-way marginal distributions of a particular characteristic such as Political Orientation changes over time, and whether these patterns are the same or not for men and women, or for young and old people. Further, are such growth curves or trends the same for two or more related characteristics, e.g., for Political party Preference and Preference Political Candidates? Or dealing (partially) with gross change: are the changes in turnover table Time 1 – Time 2 the same as the changes in turnover table Time 2 – Time 3? For the answers to all these types of questions marginal modeling is needed when using longitudinal data.

If the data come from trend studies based on repeated cross-sections, many of these questions can be answered by standard statistical techniques because the observations at different occasions are in principle independent and identically distributed. But also for trend data, sometimes marginal modeling procedures are needed, given particular research questions, e.g., when comparing the (net) changes in related characteristics. When the respondents in each of the repeated cross-sections provide information on his or her uses of alcohol (*A*), soft drugs (*S*) and hard drugs (*H*), it might be interesting to see whether the three separate one-variable growth curves for *A*, *S*, and *H* behave in the same way over time. Marginal modeling procedures are needed to test this and related hypotheses. The full table is table *TASH*, where *T* refers to the time of observation. But the information about the one way marginal distributions of *A*, *S*, and *D* is provided by the same respondents at each particular point in time.

A similar situation can occur in a single cross-sectional study. When a survey provides information about how the respondents feel about their body, i.e., about their face, eyes, legs, hips, buttocks, body build, figure, etc. Next to investigating the correlations among these aspects, it is interesting to see whether the satisfaction with particular body aspects is different from other parts, whether there is more or less variation for some parts than for others, whether these differences are the same for adult men and women, boys and girls etc. Such kinds of questions again involve dependent observations whose dependencies should be taken into account.

In a way, all the above examples essentially involve repeated measurements on the same respondents. However, clustering can occur in many other different ways. In educational research, pupils clustered in randomly chosen schools are investigated; in family research, families are randomly selected and within families husband, wife, and children; in many surveys, respondents must be seen as clustered within interviewers, and so on. Often this clustering is purposeful in the sense that the dependency is substantively interesting. But as often, and comparable to panel studies, the researchers ask questions about these data for which the dependencies are just a nuisance. Again, in such cases, marginal modeling procedures must be considered.

Many more examples could be provided but the selection presented above might be sufficient for the reader to get an idea of the usefulness of marginal modeling. More examples can be found in methodological overviews, e.g., Hagenaars, 1990, Molenberghs and Verbeke, 2005; Diggle et al., 2002; Fitzmaurice et al., 2009; Bergsma et al., 2009.

2. ML ESTIMATION AND TESTING PROCEDURES

A categorical marginal model consists of three components:

1. A collection of categorical marginal distributions
2. Coefficients defined on the marginal distributions
3. A linear model for the marginal coefficients

To take a simple example, suppose we have three categorical variables *A*, *B* and *C*, which represent measurements of the same variable at three points in time. The first component in a marginal model may consist of the bivariate marginal turnover tables *AB* and *BC*. The second component, the coefficients of interest, could be the correlation coefficients in tables *AB* and *BC*, i.e., the correlations between *A* and *B*, and between *B* and *C*. Alternatively, the second component could be the sets of marginal loglinear association coefficients (or parameters) in *AB* and *BC*, denoted by λ_{ij}^{AB} and λ_{ij}^{BC}. The third component could be the linear model for the coefficients asserting equality of correlations or marginal loglinear parameters in tables *AB* and *BC.*

The procedures and insights presented here owe much to the work of Lang and Agresti (1994), and of Grizzle Starmer and Koch (1969). Bishop et al. (1975) and especially Haber (1985) developed the first more general, but still rather restricted maximum likelihood procedures for marginal models. However, the work by Agresti and Lang, based on algorithms by Aitchison and Silvey really constituted the first very general approach towards marginal modeling using maximum likelihood procedures (Aitchison and Silvey 1958, 1960; Lang 1996a; Lang and Agresti 1994). Bergsma extended the Lang-Agresti algorithm, made it feasible for very large tables, and applied it systematically to non-loglinear models by means of the generalized exp-log notation (Bergsma, 1997). Based on this work and the work by Becker and Yang (1998), BCH made it possible to define a very general class of marginal models involving latent class models.

2.1 Practical specification of categorical marginal models

In practice, the most convenient way to specify a categorical marginal model is often as follows. Suppose $A_1, A_2, \dots, A_K$ are measurements of a categorical variable at *K* points in time, each having I categories. (Note that time is not essential here, the measurements could also be of K different items, provided all are measured on the same scale. We use time for convenience.) A new $K \times I$ marginal table TR (Time × Response) of conditional probabilities can be defined as

$$\pi_{i|t}^{R|T} = \pi_i^{A_t},$$

where $\pi_i^{A_t}$ is the marginal proportion of subjects with response *i* at time *t*. The first component of the marginal model thus consists of the table of marginal proportions $\pi_{i|t}^{R|T}$. It can be seen that the association in table *TR* relates to the differences in the marginal response distributions at the different points in time. Below, we outline three different basic approaches to modeling this association. They all involve defining association parameters

$$\theta^{TR} = f(\pi^{R|T})$$

on the marginal tables, for some appropriate function f. The θ^{TR} may be indexed as θ_{ti}^{TR} if needed.

Firstly, let us set the marginal coefficients, the second component of the model, to be the logarithms of the marginal proportions,

$$\theta_{ti}^{TR} = \log \pi_{i|t}^{R|T}.$$

A marginal loglinear model is now a linear model for the θ coefficients. For example, the independence model for table *TR* is

$$\theta_{ti}^{TR} = \lambda + \lambda_t^T + \lambda_i^R. \qquad (1)$$

Model (1) is equivalent to the marginal homogeneity model

$$\pi_i^{A_1} = \pi_i^{A_2} = \cdots = \pi_i^{A_K}$$

for $i = 1, \dots, I$. It follows that λ_t^T in (1) is constant, and can be assumed zero without any loss of generality. With θ_{ti}^{TR} the logarithms of the marginal proportions, the λ parameters in (1) are loglinear parameters. Of course, the marginal modeling approach is not restricted to the use of loglinear parameters. As an interesting alternative to the use of loglinear parameters, Ekholm et al. (1995) proposed the use of dependence ratios, defined for Table TR as

$$\theta_{ti}^{TR} = \log \frac{\pi_{i|t}^{R|T}}{\pi_i^R} .$$

Interpretational and other advantages of the dependence ratio compared to the odds ratio are listed by Ekholm (2003). With θ_{ti}^{TR} the dependence ratios instead of the log probabilities, Equation (1) still gives exactly the same marginal homogeneity model, but the λ parameters will have a different interpretation.

Model (1), whether loglinear parameters or dependence ratios are used, holds if and only if marginal homogeneity is true. A much less restrictive model has the form

$$\theta^{TR} = 0,$$

where θ^{TR} is some association coefficient for Table TR, such as the correlation coefficient or Kendall's tau. This is the second approach to modeling the association in table TR.

The third approach is a regression approach. For example, we may be interested in investigating how the (population averaged) mean response varies over time. With r_i a numerical score for category i of R, the mean response at time *t* can be denoted as

$$\theta_t^T = \sum_i r_i \pi_{ti}^{TR}.$$

Among the many familiar models for changes in means is the quadratic (marginal) regression model

$$\theta_t^T = \alpha + \beta t + \gamma t^2.$$

Although this looks like a familiar regression model, the observations at the different time points involve the same subjects, so marginal modeling techniques need to be used to find the ML estimates. The way to do this is discussed next.

2.2 Matrix formulation of Categorical Marginal Models

Before we describe the fitting procedure, we first give the matrix notation for marginal models, which is needed to implement the method on a computer. In our R-package cmm, however, many matrices have been predefined, and many models can be specified without knowledge of matrix algebra.

Denote the vector of proportions for the full table by **π.** The vector of marginal proportions of interest contains linear combinations of the elements of **π** and can be written as

$$\boldsymbol{M}'\boldsymbol{\pi},$$

where ***M*** is an appropriate matrix of zeroes and ones (for more details see BCH). We can use the generalized exp-log notation of Kritzer (1977) and BCH to represent $\boldsymbol{\theta}$, which we denote $\boldsymbol{\theta}(\boldsymbol{M}'\boldsymbol{\pi})$ to indicate its dependence on the marginal proportions. The generalized exp-log notation is very flexible, and for details we refer to BCH., but an example of the notation is as follows:

$$\boldsymbol{\theta}(\boldsymbol{M}'\boldsymbol{\pi}) = \boldsymbol{C}'\, \boldsymbol{exp}\, \boldsymbol{B}'\, \boldsymbol{log}\, \boldsymbol{A}'\boldsymbol{M}'\boldsymbol{\pi}.$$

Here, *A*, *B*, and *C* are appropriate matrices, and a wide range of coefficients, including the epsilon coefficient used in Section 3.2, or the dependence ratio of Ekholm et al. (1995), can be represented in this way, see Section 3.3.1 in BCH for details. A linear model for such a vector of coefficients can then be denoted as

$$\boldsymbol{\theta}(\boldsymbol{M}'\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}, \qquad (2)$$

for an appropriate design matrix ***X*** and a parameter vector $\boldsymbol{\beta}$.

2.3 Estimation of parameters using maximum likelihood

With $\boldsymbol{f}$ a vector of observed frequencies, the kernel of the multinomial log likelihood is given by

$$\boldsymbol{L}(\boldsymbol{\pi}|\boldsymbol{f}) = \boldsymbol{f}'\, \mathbf{log}\, \boldsymbol{\pi} - N\, \mathbf{1}'\boldsymbol{\pi}, \qquad (3)$$

where *N* is the sample size. The problem now is to find an estimator of $\boldsymbol{\beta}$ and of $\boldsymbol{\pi}$ satisfying (2), such that the multinomial likelihood (3) is maximized.

We will do this using the Lagrange multiplies technique, which is a general technique for maximizing a function subject to constraints. For this, we need to rewrite (2) in an equivalent form but without the $\boldsymbol{\beta}$ parameter. With the columns of matrix ***U*** spanning the orthogonal complement of the space spanned by the columns of **X**, we can give the equivalent representation

$$\boldsymbol{U}'\boldsymbol{\theta}(\boldsymbol{M}'\boldsymbol{\pi}) = \mathbf{0}. \qquad (4)$$

Now let $\boldsymbol{H}$ be the Jacobian of $\boldsymbol{\theta}$, i.e., the matrix with (i,j)th entry the derivative of the ith coordinate of $\boldsymbol{\theta}$ with respect to its jth argument. With $\boldsymbol{\lambda}$ a vector of Lagrange multipliers, the Lagrangian likelihood then is

$$\boldsymbol{L}(\boldsymbol{\pi}|\boldsymbol{f}) = \boldsymbol{f}' \log \boldsymbol{\pi} - N\ \mathbf{1}'\boldsymbol{\pi} - N\ \boldsymbol{\lambda}'\boldsymbol{U}'\boldsymbol{\theta}(\boldsymbol{M}'\boldsymbol{\pi}). \qquad (5)$$

Taking derivatives with respect to log $\boldsymbol{\pi}$ and equating to zero leads to the Lagrangian score equation

$$\boldsymbol{p} - \boldsymbol{\pi} - \boldsymbol{D}_{\boldsymbol{\pi}}\boldsymbol{MHU\lambda} = \mathbf{0}, \qquad (6)$$

where $\boldsymbol{p} = \boldsymbol{f}/N$ is the vector of observed cell proportions, $\boldsymbol{D}_{\boldsymbol{\pi}}$ is a diagonal matrix with the vector $\boldsymbol{\pi}$ on the main diagonal, and $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{M}'\boldsymbol{\pi})$. The maximum likelihood estimator of $\boldsymbol{\pi}$ is now a solution of (4) and (6), which can be found using a scoring type algorithm (BCH, Section 2.3.5; see also Lang and Agresti (1994), Bergsma (1997), Lang (2004)).

A main assumption for the algorithm to work is the regularity condition that there are no redundant constraints. To verify this, it is normally sufficient to check that matrix $\boldsymbol{U}$ has full column rank.

Once the ML estimates $\hat{\boldsymbol{\pi}}$ have been obtained, marginal models can be tested by means of two well-known test statistics: the likelihood ratio test statistic

$$G^2 = -2\ N \sum_i p_i \log \frac{\hat{\pi}_i}{p_i}$$

and the Pearson's chi-square test statistic

$$X^2 = N \sum_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i}\ .$$

where $p_i = f_i/N$ is the sample proportion in cell i. If the postulated model is true, these test statistics have an asymptotic chi-square distribution with degrees of freedom (*df*) equal to the number of independent constraints on the cell probabilities, which is normally equal to the column rank of $\boldsymbol{U}$.

2.4 The EM algorithm for marginal variable models with latent variables

Extending the ML algorithm so that latent variables can be included in the model is now straightforward using the EM (expectation-maximization) algorithm. The EM algorithm consists of repeated application of an E-step and an M-step, which we will explain now. First we need the concept of a complete data likelihood, by which we mean the likelihood that would have been obtained had the latent variables been observed. The complete data likelihood contains an unobserved multinomial frequency vector, which is replaced in the E-step by its expected value given the observed data and the current estimated population probabilities (Haberman, 1979; Becker and Yang 1998; BCH). For simplicity, consider a single manifest variable *A* and a latent variable *X*, and denote the estimated probability that $X = j$ given $A = i$ by $\widehat{\pi}_{j\ i}^{X|A}$. It can then be shown that expected complete data frequencies are given as

$$\hat{f}_{ij}^{AX} = f_i^A \ \hat{\pi}_{j\ i}^{X|A} .$$

The M-step now consists of maximizing the complete data likelihood, with the unobserved frequencies replaced by the $\hat{f}_{ij}^{AX}$, using the Lagrange multiplier method described in the previous subsection.

It can be shown that repeated application of the E and M steps as described above leads to a local maximum of the likelihood (Wu, 1983). However, such a local maximum is not always the global maximum, and different starting values may need to be tried to find the global maximum.

2.5 Dealing with missing data

We now outline a likelihood based method to deal with missing data. The simplest case to deal with is the case that data is *missing completely at random* (MCAR). This means that events leading to a missing observation on a particular variable are independent of both observable and unobservable variables. This case is straightforward, because only the likelihood needs to be adapted, and no extra modeling needs to be done. Alternatively, data may be *missing at random* (MAR), meaning that the missingness does not depend on the missing data itself. In this case, the missingness needs to be modeled. Fay (1986) developed a flexible approach for this, which we outline below. Finally, data may be *not missing at random* (NMAR), when being missing depends on the unseen observations themselves. In this case it may be difficult to model the missingness mechanism, and we will not discuss this case further.

Let us illustrate the MCAR case with an example of two categorical variables A and B. Suppose for some subjects, neither A nor B is observed, for some we have observations only on A, for others only on B, and for the remainder on both A and B. If missingness is MCAR, the kernel of the log likelihood is then simply the sum of the log likelihood kernels for the three groups, which gives

$$L(\boldsymbol{\pi}|\boldsymbol{f}) = \ f^* \log \pi^* \ + \sum_i f_i^A \log \pi_i^A + \sum_j f_j^B \log \pi_j^B + \sum_{i,j} f_{ij}^{AB} \log \pi_{ij}^{AB} - 4N \sum_{i,j} \pi_{ij}^{AB},$$

where f^* is the number of subjects for whom neither A nor B is observed, π^* is the corresponding expected proportion. Maximizing this likelihood subject to constraints gives the MCAR estimates. This maximization can be done using the EM algorithm, making use of the Lagrange multiplier method of Section 2.3 in the M-step for fitting the marginal constraints.

In the MAR case, the likelihood can be obtained by introducing for each variable V an additional indicator variable I_V, such that $I_V = 0$ for subjects that have a missing observation on variable V, and $I_V = 1$ otherwise. The indicator variables are observed, and $f^{I_A I_B}_{0\,0}$represents the number of subjects with missing observations on both A and B, $f^{AI_A I_B}_{i\,1\,0}$ represents the number of subjects for whom $A = i$ and which have a missing observation on B, and so forth. We obtain the following likelihood:

$$\boldsymbol{L}(\boldsymbol{\pi}|\boldsymbol{f}) = f^{I_A I_B}_{0\,0} \log \pi^{I_A I_B}_{0\,0} + \sum_i f^{AI_A I_B}_{i\,1\,0} \log \pi^{AI_A I_B}_{i\,1\,0} + \sum_j f^{BI_A I_B}_{j\,0\,1} \log \pi^{BI_A I_B}_{j\,0\,1}$$
$$+ \sum_{i,j} f^{ABI_A I_B}_{i\,j\,1\,1} \log \pi^{ABI_A I_B}_{i\,j\,1\,1} - 4N \sum_{i,j,k,l} \pi^{ABI_A I_B}_{i\,j\,k\,l}.$$

Fay’s (1986) approach involves modeling the relations among indicator and non-indicator variables by means of path models, or, more generally, loglinear models. Typically the EM algorithm will be needed to find ML estimates of cell proportions (for further details, see Fay, 1986 or Vermunt, 1997). We can then readily incorporate marginal constraints in the M-step as outlined above.

It may be wondered if problems arise when combining a marginal model with Fay’s (loglinear) constraints. Bergsma and Rudas (2002) gave general conditions on the variation independence of marginal and loglinear parameters, which guarantee the possibility of combining such marginal and loglinear constraints. (We note, however, that for some of the more complex marginal models, even without additional loglinear constraints and not involving latent variables, some difficulties may arise, e.g., in the determination of the correct number of degrees of freedom; see BCH, Section 4.5, which also gives a solution to these difficulties.)

3. ALTERNATIVE ESTIMATION METHODS

Besides maximum likelihood, there are two other popular approaches for estimating and testing marginal models: generalized estimating equations (GEE) and GSK (after Grizzle, Starmer and Koch, 1969). Below we describe advantages and disadvantages compared to each other and to the ML method. Most importantly, GEE and GSK estimates are much easier to compute than ML estimates, in particular for large numbers of variables, but they miss the flexibility and guaranteed efficiency of the ML method. For example, GEE cannot easily deal with latent variables. Standard GEE implementations allow the inclusion of continuous covariates, and we outline a loglinear model based procedure using which this can be done for marginal modeling with the ML method as well, and more efficiently. We end the section with a discussion of random coefficient models, which are also popular for modeling dependent data, but to answer different research questions.

3.1 Description of the GEE method and its relation with ML estimation

The probably most popular and widespread alternative method to ML for marginal modeling is the GEE methodology. In part to overcome some of the computational difficulties with obtaining the maximum likelihood estimates for complex marginal models, Liang and Zeger developed an extended quasi-likelihood approach called Generalized Estimating Equations (GEE) (Liang and Zeger, 1986; Diggle et al., 2002; Molenberghs and Verbeke, 2005; Lipsitz and Fitzmaurice, 2009). Recognizing that the parameter estimates in marginal models are in general consistent, even when ignoring the dependencies among the observations, the GEE approach replaces the often complex dependence structure by a much simpler one, such as independence or uniform association, and adjusts standard errors for any misspecification of the dependence using so-called sandwich estimators. A very important possibility in GEE is the use of a correlation structure which does not depend on covariates, because this allows regression parameters to be estimated consistently even if covariates are continuous. Below, we briefly describe the GEE method (more details can be found in the aforementioned literature), and in Section 3.3 we compare it with the ML method.

Like ML, the GEE approach can be used to fit marginal models of the form (2). However, standard GEE notation is slightly different, in particular, for subject i, the model of interest is written in the form

$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \qquad (7)$$

where $\boldsymbol{\mu}_i$ is some vector of (possibly marginal) coefficients, $\boldsymbol{X}_i$ is a matrix of subject specific covariates, and g is a so-called link function, which is assumed to be invertible, and which operates coordinatewise, i.e.,

$$g(\boldsymbol{\mu}_i) = \begin{pmatrix} g(\mu_{i1}) \\ \vdots \\ g(\mu_{ik}) \end{pmatrix}.$$

Let $\boldsymbol{\pi}_i$ be vector consisting of response probabilities of subject i. As shown in the appendix, if $\boldsymbol{\mu}_i = \boldsymbol{M}'\boldsymbol{\pi}_i$ is a vector of marginal proportions, and $\boldsymbol{M}$ has full column rank, the Lagrangian score equation (6) implies

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i{}'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}\,, \qquad (8)$$

where $\boldsymbol{y} = \boldsymbol{M}'\boldsymbol{p}$ is the vector of observed marginal proportions, and

$$\boldsymbol{V}_i = \boldsymbol{M}'\boldsymbol{D}_{\boldsymbol{\pi}_i}\boldsymbol{M} - \boldsymbol{M}'\boldsymbol{\pi}_i\boldsymbol{\pi}_i{}'\boldsymbol{M} \qquad (9)$$

is the covariance matrix of the observed marginal proportions for subject i.

As they stand, Equations (7) and (8) contain too many unknowns to be solved, namely the vector $\boldsymbol{\beta}$ and the off-diagonal elements of the $\boldsymbol{V}_i$. However, as noted by Liang and Zeger, if the $\boldsymbol{V}_i$ are replaced by appropriate “working” covariance matrices (which may need to be estimated), then the resulting estimator of $\boldsymbol{\beta}$ will still be consistent, even if the working covariances are wrong. Its standard error is then consistently estimated by means of the so-called sandwich estimator.

The equation (8) with $\boldsymbol{V}_i$ replaced by a working covariance matrix is called a generalized estimating equation (GEE). Normally, (8) needs to be solved using iterative methods. Note that $\boldsymbol{\mu}_i$ in (7) and (8) can represent a wide range of (marginal or non-marginal) parameters, with $\boldsymbol{y}$ the corresponding sample value.

In practice working correlations for marginal parameters are specified, from which working covariances can then be computed. Commonly used working correlation structures are *independence* (all working covariances zero), *exchangeable* (all working correlations equal, i.e., uniform association), *autoregressive* (autoregressive correlation structure), or *unstructured*. In these cases, estimation of working correlations is done by averaging conditional parameters over subjects, ensuring that working correlations are identical for all subjects (and thus do not depend on individual covariates). This ensures precise estimation of the $\boldsymbol{V}_i$ in (8) even if covariates are continuous, which in turn ensures consistent estimation of $\boldsymbol{\beta}$. The fact that estimators of the $\boldsymbol{V}_i$ may be (potentially heavily) biased does not matter for sufficiently large samples.

We can now point out the close relationship between GEE and ML estimators. Firstly, in the case that the $\boldsymbol{\mu}_i$ are vectors of marginal proportions, then, as mentioned above, the estimating equation (8) follows from the Lagrangian score equation (6), so GEE and ML are closely related. In particular, if the working correlations equal the ML estimators, then it follows that the GEE and ML estimators of $\boldsymbol{\beta}$ coincide. Secondly, in the case that the $\boldsymbol{\mu}_i$ are not vectors of marginal proportions, then (8) is not implied by (6), and GEE and ML estimators generally do not coincide. Instead, as shown in the appendix, (8) with $\boldsymbol{y}_i - \boldsymbol{\mu}_i$ replaced by a first order Taylor approximation does follow from (6). Hence in this case, for large samples, if $\boldsymbol{V}_i$ in (8) is replaced by its ML estimator, GEE and ML estimators are likely to be close together. In Section 3.3 we outline how the ML method can be used to deal with continuous covariates as well.

3.2 The GSK method

A classical approach towards marginal modeling is the GSK one after Grizzle, Starmer and Koch who wrote the first seminal article about it (Grizzle et al. 1969). It is based on Weighted Least Squares (WLS) procedures. The GSK estimator of $\boldsymbol{\beta}$ in (7) minimizes the quadratic form

$$\sum_{i=1}^{N}(g(\boldsymbol{y}_i) - \boldsymbol{X}_i\boldsymbol{\beta})'\widehat{\boldsymbol{V}}_{g(\boldsymbol{y}_i)}^{-1}(g(\boldsymbol{y}_i) - \boldsymbol{X}_i\boldsymbol{\beta}),$$

where $\widehat{\boldsymbol{V}}_{g(\boldsymbol{y}_i)}$ is the sample estimator of the covariance matrix of $g(\boldsymbol{y}_i)$ (see Agresti, 2002, Section 15.1 for further details). Now $\boldsymbol{\beta}$ can be found by taking the derivative of this expression and equating to zero, which leads to the estimating equation

$$\sum_{i=1}^{N} \boldsymbol{X}_i{}'\widehat{\boldsymbol{V}}_{g(\boldsymbol{y}_i)}^{-\mathbf{1}}(g(\boldsymbol{y}_i) - \boldsymbol{X}_i\boldsymbol{\beta}) = \mathbf{0}. \qquad (10)$$

Note the similarity of (8) and (10), in particular, if g is the identity function then GEE is in fact a generalization of GSK, allowing a broader range of choices for the covariance matrices of the $\boldsymbol{y}_i$. It can be seen that GSK requires many observations per covariate value in order to obtain a reasonable estimate $\widehat{\boldsymbol{V}}_{g(\boldsymbol{y}_i)}^{-\mathbf{1}}$, which is required to estimate $\boldsymbol{\beta}$ well (Fitzmaurice and Molenberghs, 2009). In contrast, GEE allows the assumption that the correlation matrices of the $\boldsymbol{y}_i$ are independent of i, which permits the incorporation of continuous covariates. Note however, that such assumptions are also possible in GSK, but as far as we are aware this has not been done.

The GSK estimator is asymptotically efficient, but as mentioned for small samples $\widehat{\boldsymbol{V}}_{g(\boldsymbol{y}_i)}$ may estimate the true covariance matrix $\boldsymbol{V}_{g(\boldsymbol{y}_i)}$ poorly, and a more structured working covariance may give better estimators, even if the structure is wrong, thus giving GEE an advantage. Unlike most GEE estimators, GSK estimators have a closed form and so are easier to compute. However, computation of GEE estimators generally appears to pose no major problems.

Like GEE estimates, GSK estimates are often much easier to compute than the ML estimates. Moreover, just as ML estimates, GSK estimates have desirable asymptotic properties. However, in general, for both small and large samples, ML tends to have superior properties to GSK, as is made clear in the discussion section of Berkson (1980). From a practical point of view, ML generally handles very sparse tables better and provides more reliable results for the standard errors and the test statistics. Finally, the GSK approach has not been extended to deal with latent variables, and it is not clear it will retain its (computational) advantages with such an extension.

3.3 Comparison of GEE and ML

The main advantage of ML estimation compared to GEE is its flexibility, as the likelihood can be adapted to the situation at hand. This is illustrated with the example in Section 4.2, where the (marginal) association between latent variables and observed variables is modeled, which seems impossible to do with GEE. In general, the GEE method has not been well-developed for dealing with latent variables. Furthermore, ML estimation of marginal models can readily incorporate Fay’s likelihood based method of dealing with missing data, as outlined in Section 2.5. It is true that for GEE imputation methods have been developed for dealing with missing data, but these are not as flexible as Fay's approach for modeling the missingness mechanism.

As mentioned in Section 3.1, GEE allows the assumption that the correlation matrix of the vector $\boldsymbol{y}_i$ does not depend on i, which makes the use of continuous covariates, or large numbers of categorical covariates, possible while still giving consistent estimators of $\boldsymbol{\beta}$. Without such an assumption, the $\boldsymbol{V}_i$ cannot be estimated precisely, and the $\boldsymbol{\beta}$ which solves (8) may not be consistent. The flexibility of the ML method allows a similar assumption to easily be incorporated, namely by adding loglinear constraints, in particular, that the loglinear interaction parameters for $\boldsymbol{y}_i$ do not depend on i. This assumption and the parsimonious marginal model (7) ensures that the number of free parameters in the model does not depend on the sample size N even if covariates are continuous, and so standard asymptotic theory applies, ensuring the ML estimator of $\boldsymbol{\beta}$ is consistent (e.g., Agresti, 2013, Chapter 16). Hence this method provides a model-based analogue of unstructured working correlations in GEE, where the assumption that correlations among responses do not depend on covariates is replaced by the assumption that loglinear interaction parameters for responses do not depend on covariates. Further assumptions in the ML method can be made to mimic structured working correlations in GEE. Important to note here is that $\boldsymbol{\beta}$ is orthogonal to the loglinear parameters which are set to zero, because the observations y_{ik} are sufficient statistics for the proposed loglinear model, and $\boldsymbol{\beta}$ is a function of the $Ey_{ik} = \mu_{ik}$ (see Lang, 1996b). This ensures asymptotic efficiency in estimating $\boldsymbol{\beta}$ is retained and the sandwich correction does not need to be applied to estimated standard errors, even if the loglinear model is wrong. This asymptotic efficiency is not shared by GEE estimators, because conditional correlations are not orthogonal to marginal parameters, and so in this aspect ML estimators have an important advantage compared to GEE estimators.

However, GEE does have a major advantage compared to ML, namely its computational simplicity, allowing it to deal with rather large numbers of variables. For the ML method, the computational complexity increases exponentially with the number of variables, so no matter how fast computers will become in the future, it will always be the case that only a limited number of variables can be dealt with. Notwithstanding this, ML has broader scope than is commonly thought, and currently we can deal with about a million cells in a contingency table, which amounts to 20 dichotomous variables, 13 trichotomous ones, or 8 variables with 5 categories each.

As mentioned above, ML estimators are guaranteed to be asymptotically efficient, whereas GEE estimators are only so if the working covariance matrices are consistent, which is unlikely in practice. Nevertheless, it has been noted that in many practical situations GEE's efficiency loss is not big, and this has been our experience as well in simulations we have performed. We also found that for commonly used working covariances, GEE often, but not always, performs well compared to ML even if these working covariances are far off from the truth. The following simplified examples illustrate when GEE does and when it does not perform well. Consider the model of marginal homogeneity for the univariate margins of a $2 \times 2 \times 2$ table ABC, i.e., the model

$$\pi_t^A = \pi_t^B = \pi_t^C = \begin{cases} \beta & t = 1 \\ 1-\beta & t = 2 \end{cases}.$$

The GEE estimator of β, assuming an independence working correlation matrix, is simply the average of the marginal observed proportions, i.e.,

$$\tilde{\beta} = \frac{p_1^A + p_1^B + p_1^C}{3}.$$

The ML estimator $\hat{\beta}$ does not in general have a closed form expression for this model. Let us now compare the efficiency of $\tilde{\beta}$ and $\hat{\beta}$ in two extreme situations: firstly that A, B and C are all perfectly positively correlated, and secondly that A and B are perfectly positively correlated, and both perfectly negatively correlated with C. In the first situation, it can be shown that the ML and GEE estimators coincide, so the fact that the working correlation is far off does not negatively affect the estimator compared to ML. In the second situation, it can be shown that the ML estimator has zero variance, while the GEE estimator has variance $1/(12N)$. This is a pattern we found generally, using independence, exchangeable, or autoregressive working correlations: if the correlations among the marginal distributions do not differ too much, then GEE using standard working correlations and ML estimators have similar efficiency, while if there are large differences in marginal correlations, ML can significantly outperform GEE.

Unlike GEE, the likelihood method gives overall goodness-of-fit statistics, such as the likelihood ratio test or Pearson chi-squared test. Instead, for GEE Wald type tests are commonly used, e.g., to test a linear regression line against the alternative of a quadratic regression line. A summary of other methods can be found in Lipsitz and Fitzmaurice (2009, Section 3.5).

We finally note there exist some misconceptions about the drawbacks of likelihood based methods compared to GEE. One common perception appears to be that likelihood based methods require a parameterization involving both marginal and higher order interaction parameters (e.g., Fitzmaurice and Molenberghs, 2009, p.14). But such a parameterization is clearly not necessary if the Lagrange multiplier technique outlined in Section 3 is used. A broad family of parameterizations is given by Bergsma and Rudas, 2002, but these are useful for modeling purposes and especially for determining the properties of models, and are unnecessary, and could even be cumbersome, for ML algorithms.

3.4 Random coefficient models

At least among social scientists, random coefficient models, also denoted as conditional, cluster specific, or subject-specific models, may well be the standard way of handling dependent observations (Agresti 2002; Agresti, 2013; Raudenbusch and Bryk 2003; Molenberghs and Verbeke 2005). However, marginal models (sometimes also called population averaged models) and random effect models are generally used to answer different substantive research questions. They lead to different estimators which may also have very different substantive interpretations. Imagine a growth curve study where the dependent variable is being Conservative or not, and imagine that the effects of age are such that for each additional year there is a linear increase in the probability of being Conservative of .005 (e.g., at Age 18: .300; at age 19: .305; at age 20: .310, etc.). If these estimates had been obtained from a trend study in which at each successive year (age) a new sample was drawn from the same birth cohort, the interpretation of the age effect would run as follows: a randomly chosen person from the age group 18 (the average cohort member in the population at age 18) has a probability of being Conservative that is 10 x .005 = .05 less than the probability that a randomly chosen person from this cohort at age 28 has of being conservative. If the data had come from a longitudinal study, following the same random sample from this birth cohort over time, and the estimates were obtained by marginal modeling, the interpretation would be exactly the same as for repeated independent samples. However, if in the longitudinal study a random coefficient model was applied to obtain the estimates, the interpretation would have been different because one conditions on the unobserved characteristics of the individuals: a randomly chosen person from age group 18 has a probability of being Conservative that is 10 x .005 = .05 less than for a randomly chosen person from this cohort at age 28, provided that the two individuals have the same unobserved characteristics. The one interpretation is not to be automatically preferred above the other. It obviously depends on the nature of the research question whether the marginal or the conditional approach is more adequate.

Typically in the random coefficient literature, research questions about marginal distributions are handled by integrating out random coefficients. However, this may be computationally cumbersome, and the random coefficient models typically make needlessly restrictive and often unverifiable assumptions about the (nuisance) dependence structure. The marginal modeling approach advocated in this paper, in which assumptions about these dependencies do not need to be made, is much more flexible and realistic in this respect.

4. EXAMPLES

In this section two examples of marginal analyses on categorical data are presented. In the first example the stability of the association between two categorical variables over time is investigated on data collected in a complex rotating panel design. The second example illustrates how marginal models can be extended to include latent variables, and how the interaction between the latent and manifest variables can be defined using a non-loglinear approach using the *epsilon* (ε) association coefficient instead of the better known loglinear two-variable interaction terms.

4.1 Analyzing data from a complex rotating designs

In Chapter 4 of BCH the authors give due attention to the applicability of marginal models to longitudinal data collected in either a repeated cross-section or a panel study. In large scale social surveys often more complex designs are used, such as, for instance, a rotating panel survey, in which several subsamples are involved and each subsample is observed at multiple time points before being replaced by a new subsample. These designs, which also are referred to as accelerated longitudinal designs, combine the advantages of both panel and cross-sectional surveys.

The Italian Continuous Labor Force Survey, supervised by *Istat*, the Italian National Institute of Statistics, collects data in a 2-2-2 rotating design on the labor market participation of respondents from the non-institutional Italian population. Each rotation group enters the study at a particular quarter of the year and is first observed for two consecutive quarters, then left out of the study for the next two quarters, before being interviewed again for two final consecutive quarters. In this way, seven different rotation groups span a period of three years. Table 3 shows the details of this rotating design covering the period 2004-2006

Table 3. Structure of the 2-2-2 rotation design from the Italian Continuous Labor Force Survey covering the period 2004-2006.

	2004				2005				2006			
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
RG1	+	+	-	-	+	+	-	-	-	-	-	-
RG2	-	+	+	-	-	+	+	-	-	-	-	-
RG3	-	-	+	+	-	-	+	+	-	-	-	-
RG4	-	-	-	+	+	-	-	+	+	-	-	-
RG5	-	-	-	-	+	+	-	-	+	+	-	-
RG6	-	-	-	-	-	+	+	-	-	+	+	-
RG7	-	-	-	-	-	-	+	+	-	-	+	+

Data collected in this rotation design are partially dependent and partially independent. An assessment of changes between the first and the last quarter, for instance, only requires the comparison of independent data from the first and the seventh rotation group. On the other hand, an assessment of the net changes between the sixth and seventh quarters is partially based on independent data from the various rotation groups, but since the second and the sixth rotation group have observations at both quarters, part of the comparison will include dependent data as well.

In the Continuous Labor Force Survey several measures for labor market participation of individuals are defined. First, each respondent is classified as employed, unemployed, or out of the labor force according to the definition of the International Labor Office (ILO). This classification is based on the respondent's answers to several questions regarding his recent work situation. Second, a self-perception (SP) indicator is obtained by asking each respondent to classify himself as being employed, unemployed, or out of the labor force. In what follows, both the ILO and the SP measure will be treated as categorical variables with three response categories:

1 = employed, 2 = unemployed, 3 = out of the labor force.

Only respondents with complete data on both measures at the four measurement occasions were retained in the analysis. The number of respondents in each rotation group then varied around 27,000. The total number of respondents was 194,549.

As a preliminary step in the analysis, the 4 × 3 × 3 Occasion × ILO × SP table was defined for each rotation group. Merging those seven tables in the appropriate way allows the construction of the 12 × 3 × 3 table Quarter (Q) × ILO (I) × SP (S), which is shown in Table 4 (QIS).

Table 4. The Quarter x ILO x SP table (QIS)

	ILO=1			ILO=2			ILO=3		
	SP=1	SP=2	SP=3	SP=1	SP=2	SP=3	SP=1	SP=2	SP=3
Q1	11342	188	405	2	887	187	36	837	14556
Q2	23282	307	632	10	1632	329	40	1579	28904
Q3	23544	379	725	9	1563	298	54	1795	29131
Q4	23994	342	649	5	1704	303	56	1613	29186
Q5	34956	433	711	14	2597	385	73	2622	43188
Q6	46285	450	761	13	3004	507	86	3138	57000
Q7	45294	461	881	13	2685	439	91	3780	55926
Q8	34713	404	555	6	2380	339	70	2491	42347
Q9	23676	338	471	3	1642	268	52	1780	28309
Q10	23003	204	383	5	1278	224	52	1594	27786
Q11	21738	185	323	7	1055	148	39	1850	26727
Q12	10735	88	147	0	558	97	12	746	13070

It is important to realize that the twelve different 3 x 3 tables reported in Table 4 are not based on completely independent observations, since each respondent is interviewed four times, and hence contributes to each of the tables of the quarters in which he was interviewed.

In the context of a study of the equivalence of the two measures, one could first look at their association. A quick glimpse at each of the 12 separate ILO x SP tables shows that both variables are strongly associated at each quarter, and one can then ask whether this association remains stable over time. Testing the hypothesis of a constant association over time amounts to the same as testing the fit of the no-three-variable interaction loglinear model [*QI, QS, IS*] to Table *QIS*:

$$\ln \pi_{qis}^{QIS} = \lambda + \lambda_q^Q + \lambda_i^I + \lambda_s^S + \lambda_{qi}^{QI} + \lambda_{qs}^{QS} + \lambda_{is}^{IS}.$$

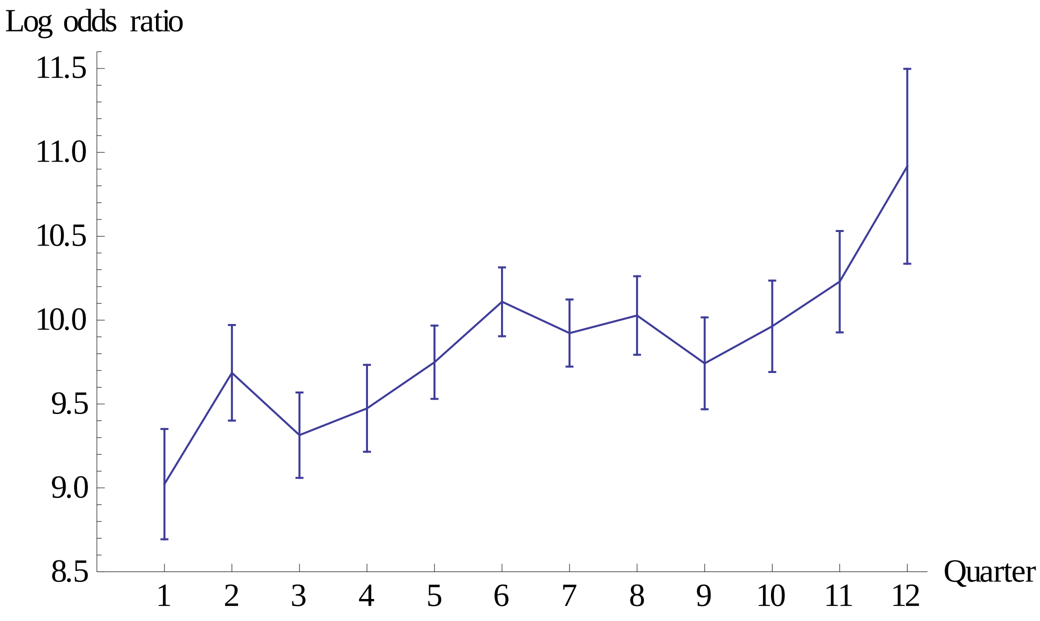
This model represents the hypothesis that the association between *I* and *S* does not depend on *Q*, although *Q* may have main effects on *I* and *S*. Correctly taking into account the observational dependencies among the data, analysis according to this model yields a test statistic $G^2 = 160.207$ $(X^2 = 164.024)$ with 44 degrees of freedom, which leads to a clear rejection of the proposed model. If the same model is tested without taking the dependencies into account by treating the 12 subtables as being based on independent samples, the test statistic becomes $G^2 = 166.626$ $(X^2 = 173.61)$. In this example the “wrong” test statistic is larger than the “correct” statistic although both variables are strongly positively associated. Both analyses lead to the same qualitative conclusion, but this is of course due to the very large samples involved in the analyses. That the difference between the two test statistics is not very impressive may be explained by noting that the data come from seven independent rotation groups so that the dependency in the data is not extreme.

In order to get an idea of what changes in association take place over time, one could look at changes in various local or global odds ratios. Here attention will be restricted to the global log odds ratio

$$\log\left(\frac{\Pr(ILO = 1, SP = 1)\Pr(ILO \neq 1, SP \neq 1)}{\Pr(ILO = 1, SP \neq 1)\Pr(ILO \neq 1, SP = 1)}\right),$$

which is the log odds ratio obtained by dichotomizing both variables with response = 1 (employed) versus response either = 2 (unemployed) or = 3 (out of the labor force). Hence, the original response categories 2 and 3 are collapsed into a single category. All the twelve log odds ratios proved to be larger than 9 and, moreover, to exhibit a clear increase over time, as is shown in Figure 1.

Figure 1. Log odds ratio for ILO and SP as function of Quarter. The vertical bars represent 95% confidence intervals.



The results of an orthonormal trend analysis with components up to the quartic are given inTable 5.

Table 5. Orthonormal trend analysis on log odds ratios

	B	SE	Z
Linear	1.35	0.20	6.86
Quadratic	0.11	0.19	0.59
Cubic	0.51	0.18	2.76
Quartic	0.36	0.17	2.19

These results show that there is a strong linear component in the overall trend for this log odds ratio, and, although the quadratic component is not significant, the cubic and quartic both are.

4.2 Non-loglinear Latent Class Models

There exists a classical data set on Party and Candidate Preference, viz. Lazarsfeld's 1940 data on Party and Candidate Preference in Erie County, Ohio (Lazarsfeld 1972, p. 392). This data set is presented in Table 6. 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.

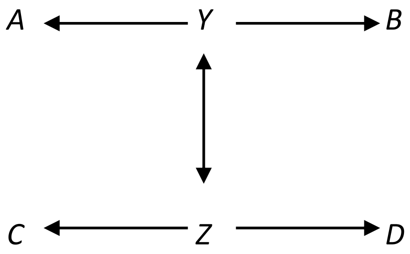
Table 6. Party Preference (PP) and Presidential Candidate Preference (CP); Erie County Ohio, 1940; t_1 – August, t_2 – October

		C. CP – t_1	1. Dem.	1. Dem.	2. Rep.	2. Rep.
A.PP-t_1	B-PP-t_2	D.CP – t_2	1. Dem.	2. Rep.	1. Dem.	2. Rep.
1. Dem.	1. Dem.		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

Hagenaars (1993) fitted several latent class models to the data in Table 6. A graphical representation of the comparatively best fitting latent class model is depicted in Figure 2, in which *A* through *D* refer to the variables in Table 6 and *Y* and *Z* are two dichotomous latent variables with *Y* representing latent party preference and *Z* latent candidate preference. The model depicted here assumes that there is no change over time in both latent variables, but that each latent variable is measured twice by an unreliable indicator variable.

Figure 2 : Latent class model for data in Table 6.



The basic latent class analysis (LCA) equation for the model in Figure 2 can be written as

$$\pi_{yzabcd}^{YZABCD} = \pi_{yz}^{YZ}\pi_{abcd\ yz}^{ABCD|YZ} = \pi_{yz}^{YZ}\pi_{a\,y}^{A|Y}\pi_{b\,y}^{B|Y}\pi_{c\,z}^{C|Z}\pi_{d\,z}^{D|Z}, \qquad (11)$$

where π_{yz}^{YZ} represents the joint probability of scoring (*y,z*) on *YZ*, $\pi_{a\,y}^{A|Y}$ the conditional response probability of scoring *A=a*, given *Y=y*, and the other symbols have obvious analogous meanings. The first part of Equation (11) ($\pi_{yzabcd}^{YZABCD} = \pi_{yz}^{YZ}\pi_{abcd\ yz}^{ABCD|YZ}$) is a tautology and by definition true, as it follows from basic rules of probability calculus. However, under the assumption of local independence, 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 Equation (11), *i.e.*, as $\pi^{YZ}_{yz}\pi^{A|Y}_{ay}\pi^{B|Y}_{by}\pi^{C|Z}_{cz}\pi^{D|Z}_{dz}$. Note that here it is further assumed that latent variable *Y* has only an effect on *A* and *B*, whereas latent variable *Z* has only an effect on *C* and *D*.

The model in Figure 2 can equivalently be represented as loglinear model *[YZ,YA,YB,ZC,ZD],* using the usual short hand notation for denoting hierarchical loglinear models, written out in full as

$$\ln \pi^{YZABCD}_{yzabcd} = \lambda + \lambda^{Y}_{y} + \lambda^{Z}_{z} + \lambda^{A}_{a} + \lambda^{B}_{b} + \lambda^{C}_{c} + \lambda^{D}_{d} + \lambda^{YZ}_{yz} + \lambda^{YA}_{ya} + \lambda^{YB}_{yb} + \lambda^{ZC}_{zc} + \lambda^{ZD}_{zd}\,. \qquad (12)$$

This model fits the data in Table 6 well with G^2 = 7.32, df = 4, p = .120 (X^2 = 11.53). Because this example concerns different kinds of restrictions on the parameters of Equation (11) and (12), they are given here in Table 7.

Table 7. Estimates of Parameters in Equations (11) and (12) applied to the data in Table 6

Y=y	*Z=z*	$\hat{\pi}^{YZ}_{yz}$	$\hat{\pi}^{A\|Y}_{1y}$	$\hat{\pi}^{A\|Y}_{2y}$	$\hat{\pi}^{B\|Y}_{1y}$	$\hat{\pi}^{B\|Y}_{2y}$	$\hat{\pi}^{C\|Z}_{1z}$	$\hat{\pi}^{C\|Z}_{2z}$	$\hat{\pi}^{D\|Z}_{1z}$	$\hat{\pi}^{D\|Z}_{2z}$
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
		$\hat{\lambda}^{YZ}_{11} = .874$		$\hat{\lambda}^{YA}_{11} = 1.916$	$\hat{\lambda}^{YB}_{11} = 2.567$		$\hat{\lambda}^{ZC}_{11} = 1.046$		$\hat{\lambda}^{ZD}_{11} = 4.967$	
		$\hat{\varepsilon}^{Z\|Y}_{11} = .703$		$\hat{\varepsilon}^{A\|Y}_{11} = .952$	$\hat{\varepsilon}^{B\|Y}_{11} = .987$		$\hat{\varepsilon}^{C\|Z}_{11} = .772$		$\hat{\varepsilon}^{D\|Z}_{11} = .986$	
				(s.e. = .024)	(s.e. = .042)		(s.e. = .016)		(s.e. = .022)	

Latent class outcomes always contain a lot of detailed and interesting information, which will be largely ignored here. The focus will be on 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 2002, 2010).

The loglinear parameterization of the latent class model in Equation (12) is identical to its general formulation in Equation (11) in the sense that they yield the same estimated probabilities for the full table *ABCDYZ* if no further restrictions are imposed on the parameters (except for the usual identifying restrictions). Therefore, the strength and direction of the relationships between the variables can be expressed by means of the two-variable loglinear parameters from Equation (12), which can be computed on the basis of the conditional response probabilities in Equation (11). The pertinent $\hat{\lambda}$ estimates are reported at the second-to-last row of Table 7. According to these loglinear association coefficients, manifest variable *D* is the most reliable indicator, followed by *B, A*, and *C*. Note however that the very large size of the effect of *Z* on *D* is a consequence of the fact that Table *ZD* contains an almost empty cell with $\hat{\pi}_{1\,2}^{D|Z}$ < 0.001.

However, expressing the directions and strength of the relationships among the variables in terms of the loglinear parameters and odds ratios, in other words, parameterizing the basic latent class as a loglinear model, is in a way arbitrary. For example, researchers may 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 odds ratios. Coefficient ε is a measure of the strength of the effect of an independent variable *X* on a dependent variable *Y*. When both variables are dichotomous with scores 0 and 1, coefficient ε is defined as a difference of two conditional probabilities:

$$\varepsilon = P(Y = 1|X = 1) - P(Y = 1|X = 0).$$

This coefficient is actually the regression coefficient with *Y* regressed on *X*. For example, the effect of Y on A can is estimated as follows, using the estimated conditional response probabilities in Table 7:

$$\hat{\varepsilon}_{11}^{A|Y} = \hat{\pi}_{11}^{A|Y} - \hat{\pi}_{12}^{A|Y} = .965 - .013 = .952\,.$$

In the present context these values of ε can be interpreted as reliabilities, since they indicate how strongly each observed indicator is related to the latent variable it should measure. The estimated ‘reliabilities’ in terms of ε are presented in the last row of Table 7. 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.

As long as no further restrictions are imposed on the parameters, it is largely a matter of the researcher’s reasoned preferences whether to express the basic latent class model as a multiplicative/loglinear model with the *λ*-parameters or as a basically additive model and use the ε’s, as long as both formulations lead to the same estimated probabilities for the joint table. However, the explicit choice of an appropriate parameterization becomes more urgent and even necessary if (additional) restrictions are imposed on the LCA model that lead to different implications for the data, essentially concerning restrictions 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 of the probabilities 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. 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 when the pertinent ε’s are set equal.

Imposing the 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 (Vermunt 1997b), MPLUS (Muthen and Muthen 2006), or Latent Gold (Vermunt and Magidson 2005).

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 so that the standard (Goodman/Haberman) routines can no longer be used. However, an appropriate ML estimation procedure is provided by the marginal modeling approach.

Some of the important outcomes applying the standard procedure where possible, as well as the marginal modeling approach are as follows. 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 < .001). 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: the strict equalities 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 that did fit was the model in which change was allowed in the reliability of candidate preference but the reliabilities of party preference were assumed not to change:

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

The test outcomes are: G^2 = 8.45, df = 5, p = .133. The reliabilities were estimated as

$$\hat{\varepsilon}_{1\,1}^{A|Y} = \hat{\varepsilon}_{1\,1}^{B|Y} = .969 \quad (s.e.\ 0.012)$$
$$\hat{\varepsilon}_{1\,1}^{C|Z} = .774 \quad (s.e.\ 0.042)$$
$$\hat{\varepsilon}_{1\,1}^{D|Z} = .981 \quad (s.e.\ 0.023).$$

Different conclusions can and sometimes will be reached when different parameterizations are applied. The marginal modeling approach offers the researcher more possibilities to choose from and in this way more chance of performing analyses that are closer to one's research questions.

5. CONCLUSION

Marginal modeling of categorical data provides very important extensions of categorical data analysis techniques for situations where the data are dependent, and the dependencies are not of primary interest. Dependent, or clustered, data occur a lot in practice, so this extension is important. Moreover, the methodology used for marginal modeling can be used outside the clustering context for other nonstandard situations, for example, to estimate correlation or association models that fall outside the exponential family, as shown by the second example in the previous section.

The maximum likelihood methodology of this paper is rather flexible and efficient in handling large tables, but still some work needs to be done to make it suitable for very large problems, say, marginal analysis for longitudinal studies with at least, say, 10 to 20 waves, depending on the number of categories per variable. The discussion here was limited to marginal models for categorical data. In fact, marginal models for continuous data may be equally interesting. But these have been around for a long time, not always under the name of marginal models but, for example, under the disguise of MANOVA and the like. BCH discuss several of these models (BCH, Section 7.1).

Finally, and perhaps most importantly, estimation procedures cannot be used in practice unless appropriate computer programs are offered. Bergsma and Van der Ark have developed a *Mathematica* and an *R* package version to estimate marginal models (BCH; Bergsma and Van der Ark 2009). More information can be found on the website www.cmm.st developed and maintained by Bergsma.

ACKNOWLEDGEMENTS

We are grateful to two anonymous reviewers whose comments have helped to substantially improve the paper.

APPENDIX

We will show how the GEE estimating equation (8) can be derived from the multinomial score equation (6). In particular, if $\boldsymbol{\mu}$ is a vector of marginal proportions, then (8) follows from (6), while otherwise (8) follows from a first order Taylor approximation to (6).

Denote the derivative of g by $\dot{g}$ and let $\boldsymbol{D}_i$ be the diagonal matrix with the derivative vector $\dot{g}(\boldsymbol{\mu}_i)$ on the main diagonal. Then by standard analysis,

$$\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} = \boldsymbol{X}_i'\boldsymbol{D}_i^{-1}.$$

Hence, (8) reduces to

$$\sum_{i=1}^{N} \boldsymbol{X}_i{}'\boldsymbol{D}_i^{-1}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}. \qquad (13)$$

Writing

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_N \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix},$$

and

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}_N \end{pmatrix}, \qquad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{V}_N \end{pmatrix},$$

we can write (13) as

$$\boldsymbol{X}'\boldsymbol{D}_{\dot{g}}^{-1}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}. \qquad (14)$$

Suppose the columns of $\boldsymbol{U}$ span the orthogonal complement of the columns of $\boldsymbol{X}$. Then (14) holds if and only if there exists a $\boldsymbol{\lambda}$ such that

$$\boldsymbol{D}_{\dot{g}}^{-1}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{U\lambda} = \boldsymbol{0}.$$

Premultiplying both sides by $\boldsymbol{D}_{\dot{g}}\,\boldsymbol{V}$ shows this is equivalent to

$$\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{V}\boldsymbol{D}_{\dot{g}}\boldsymbol{U\lambda} = \boldsymbol{0}. \quad (15)$$

Write

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}_1 \\ \vdots \\ \boldsymbol{\pi}_N \end{pmatrix},$$

where $\boldsymbol{\pi}_i$ is the probability vector for subject i, i.e., each subject has its own probability vector. First suppose $\boldsymbol{\mu} = \boldsymbol{M'\pi}$ is a vector of marginal probabilities and assume $g(\boldsymbol{\mu})$ is a so-called homogeneous function of $\boldsymbol{\pi}$, which is usually the case in practice (see Section 3.3.3 in BCH for details). Then $\boldsymbol{V}$ is given by

$$\boldsymbol{V} = \boldsymbol{M}'\boldsymbol{D}_{\pi}\boldsymbol{M} - \boldsymbol{M}'\boldsymbol{\pi}\boldsymbol{\pi}'\boldsymbol{M},$$

and homogeneity implies that (15) is equivalent to

$$\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{M}'\boldsymbol{D}_{\pi}\boldsymbol{M}\boldsymbol{D}_{\dot{g}}\boldsymbol{U\lambda} = \boldsymbol{0}.$$

But this equation follows from the Lagrangian score equation (6) by premultiplying both sides by $\boldsymbol{M'}$. Hence, we have shown that the estimating equation (8) is *implied* by the multinomial score equation (6).

If, on the other hand, $\boldsymbol{\mu}$ is a nonlinear function of $\boldsymbol{\pi}$, then (15) does not follow from (6). However, the two equations are closely related, which can be seen as follows. Let $\boldsymbol{S}$ be the Jacobian of $\boldsymbol{\mu}$. Then, using the delta method, the asymptotic covariance matrix of $\boldsymbol{y}$ is found to be

$$\boldsymbol{V} = \boldsymbol{S}'\boldsymbol{M}'\boldsymbol{D}_{\pi}\boldsymbol{M}\boldsymbol{S} - \boldsymbol{S}'\boldsymbol{M}'\boldsymbol{\pi}\boldsymbol{\pi}'\boldsymbol{M}\boldsymbol{S}, \quad (16)$$

and again under homogeneity of $g(\boldsymbol{\mu})$, (15) with $\boldsymbol{V}$ replaced by (16) is equivalent to

$$\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{S}'\boldsymbol{M}'^{\boldsymbol{D}_{\pi}}\boldsymbol{M}\boldsymbol{S}\boldsymbol{D}_{\dot{g}}\boldsymbol{U}\boldsymbol{\lambda} = \boldsymbol{0}. \quad (17)$$

A Taylor expansion of the difference $\boldsymbol{y} - \boldsymbol{\mu}$ is given as

$$\boldsymbol{y} - \boldsymbol{\mu} = \boldsymbol{S}'\boldsymbol{M}'(\boldsymbol{p} - \boldsymbol{\pi}) + O(||\boldsymbol{p} - \boldsymbol{\pi}||^2).$$

Since $\boldsymbol{p}$ approaches $\boldsymbol{\pi}$ as the sample size goes to infinity, for large samples the remainder term $O(||\boldsymbol{p} - \boldsymbol{\pi}||^2)$ will then become negligible compared to the other terms. Replacing $\boldsymbol{y} - \boldsymbol{\mu}$ in (17) by its first order Taylor approximation $\boldsymbol{S}'\boldsymbol{M}'(\boldsymbol{p} - \boldsymbol{\pi})$ gives

$$\boldsymbol{S}'\boldsymbol{M}'(\boldsymbol{p} - \boldsymbol{\pi}) - \boldsymbol{S}'\boldsymbol{M}'\boldsymbol{D}_{\pi}\boldsymbol{M}\boldsymbol{S}\boldsymbol{D}_{\dot{g}}\boldsymbol{U}\boldsymbol{\lambda} = \boldsymbol{0}.$$

This equation follows from the Lagrangian score equation (6) by premultiplying both sides by $\boldsymbol{S}'\boldsymbol{M}'$.

REFERENCES

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

Aitchison, J. and Silvey. S.D. 1960 “Maximum-likelihood estimation procedures and associated tests of significance.” *Journal of the Royal Statistical Society, Series B*: 154-171.

Agresti, A. 2002. *Categorical Data Analysis (Second Edition).* Hoboken, Wiley

Agresti, A. 2013. *Categorical Data Analysis (Third Edition).* Hoboken, Wiley

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.

Bartolucci, F. and A. Forcina 2002. “Extended RC association models allowing for order restrictions and marginal modelling.” *Journal of the American Statistical Association*, 97: 1192-1199.

Becker, M.P. and Yang, I. 1998. “Latent class marginal models for cross-classifications of counts.” In A.E. Raftery (Ed.), *Sociological Methodology, 1998*: 293-326. Oxford: Blackwell

Bergsma, W. P. 1997. *Marginal Models for Categorical Data*. Tilburg: Tilburg University Press.

[BCH] Bergsma, W., Croon, M., and Hagenaars, J.A. 2009. *Marginal models for dependent, clustered, and longitudinal categorical data.* New York: Springer.

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

Bergsma, W. P., & Van der Ark, L. A. 2012. *cmm: Categorical marginal models. R package version 0.4.*

Berkson, J. 1980. "Minimum chi-square, not maximimum likelihood!" *Annals of Statistics*, 8: 457-487

Bishop, Y.V.V., Fienberg, S.E. and Holland, P.W. 1975. *Discrete Multivariate Analysis*. Cambridge: MIT Press

Caussinus, H. 1966. "Contribution à l'analyse statistique des tableaux de correlation." *Annales de la Faculté des Sciences de l'Université de Toulouse,* 29: 77-182

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

Diggle, P.J., Heagerty, P.J., Liang, K.Y., and Zeger, S.L. 2002. *Analysis of Longitudinal Data.* Oxford: Oxford University Press

Duncan, O.D. 1979. "Testing key hypotheses in panel analysis." In K.F. Schuessler (Ed.), *Sociological Methodology, 1980*: 279-289. San Francisco: Jossey-Bass

Duncan, O.D. 1981. "Two faces of panel analysis: parallels with comparative cross-sectional analysis and time-lagged association." In Leinhardt (Ed.), *Sociological Methodology, 1981*: 281-318. San Francisco: Jossey-Bass

Edwards, D. 2000. *Introduction to graphical modelling*. New York: Springer.

Ekholm, A., Smith, P. W. F. & McDonald, J. W. 1995. "Marginal regression analysis of a multivariate binary response." *Biometrika*, 82, 847-854

Ekholm, A. 2003. "Comparing the odds and the dependence ratios." In Hoglund, Jantti, Rosenqvist (eds), Statistics, Econometrics and Society: Essays in honour of Leif Nordberg, pages 13-25. Helsinki: Finland

Fay, R. E. 1986. "Causal models for patterns of nonresponse." *Journal of the American Statistical Association*, *81*, 354-365.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.) 2009. *Longitudinal data analysis: A handbook of modern statistical methods*. Chapman & Hall/CRC.

Fitzmaurice, G. and Molenberghs, G. 2009. "Advances in longitudinal data analysis: *An historical perspective*." In: Fitzmaurice, Davidian, Verbeke, Molenberghs (Eds.), *Longitudinal Data Analysis*, 43-78.

Goodman, L. A. 1973. "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. 1974." Exploratory latent structure analysis using both identifiable and unidentifiable models." *Biometrika*, *61*, 215-231.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. 1969. "Analysis of categorical data by linear models." *Biometrics,* 25:489-504

Haber, M. 1985. "Maximum likelihood methods for linear and loglinear models in categorical data." *Computational Statistics and Data Analysis,* 3:1-10

Haberman, S.J. 1979. *Analysis of Qualitative Data, Volume 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*: 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., & McCutcheon, A. L. 2002. *Applied latent class analysis.* Cambridge: Cambridge University Press

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*: 234-286. Cambridge: Cambridge University Press.

Hagenaars J.A. 2010. “Loglinear latent variable models for longitudinal categorical data.” In Van Montfort K., Oud, H., and Satorra A. (Eds.), *Longitudinal Research with Latent Variables*: 1-36. Berlin: Springer

Hagenaars, J.A., Bergsma W., Croon, M. 2013 (forthcoming). “Nonloglinear marginal latent class models.” In G.R. Hancock and G.B. Macready (Eds.). *Advances in latent Class analysis: A Festschrift in Honor of C. Mitchell Dayton.* Charlotte,NC: Information Age Publishing

Koch, G.G., Landis, J.R., Freemann, D.H., and Lehnen, R.G. 1977. “A general methodology for the analysis of experiments with repeated measurements of categorical data.” *Biometrics*, 33: 133-158.

Kritzer, H.M. 1977. “Analyzing measures of association derived from contingency tables.” *Sociological Methods and Research*, 5: 35-50

Lang, J. B. 1996a. “Maximum likelihood methods for a generalized class of log-linear models.” *The Annals of Statistics*, 24: 726-752.

Lang, J. B. 1996b. “On the partitioning of goodness-of-fit statistics for multivariate categorical response models.” *Journal of the American Statistical Association*, 91: 1017-1023.

Lang, J.B. 2004. “Multinomial-Poisson homogeneous models for contingency tables.” Annals of Statistics, 32: 340-383.

Lang, J.B. and Agresti, A. 1994. “Simultaneously modeling the joint and marginal distributions of multivariate categorical responses.” *Journal of the American Statistical Association,* 89: 625-632.

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*: 388-398. New York: Free Press.

Lazarsfeld, P. F., & Henry, N. W. 1968. *Latent Structure Analysis*. Boston, MA: Houghton Mifflin.

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

Liang, K.Y and Zeger, S.L. 1986. “Longitudinal data analysis using generalized linear models.” *Biometrika,* 73: 13-22.

Lipsitz, S., & Fitzmaurice, G. 2009. “Generalized estimating equations for longitudinal data analysis.” In: Fitzmaurice, Davidian, Verbeke, Molenberghs (eds), *Longitudinal Data Analysis*, 43-78.

Lipsitz, S. R., Laird, N. M., & Harrington, D. P. 1991. “Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association.” *Biometrika*, 78: 153-160.

Molenberghs, G. and Verbeke, G. 2005. *Models for Discrete Longitudinal Data*. New York: Springer

Muthén, L. K., & Muthén, B. O. 1998. *Mplus: Statistical analysis with latent variables. (User’s guide sixth edition)*. Los Angeles, CA: Muthén and Muthén.

Raudenbush, S.W. and Bryk, A.S. 2003. *Hierarchical Linear Models: Applications and Data analysis methods.* Thousand Oaks: Sage.

Skrondal, A., and Rabe-Hesketh, S. 2004. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman and Hall.

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., Rodrigo, M.F., and Ato-Garcia, M. 2001. “Modeling joint and marginal distributions in the analysis of categorical panel data.” *Sociological Methods and Research*, 30: 170-196.

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

Wu, C. F. 1983. “On the convergence properties of the EM algorithm.” *The Annals of Statistics*, 11: 95-103.