Fast estimation algorithm for likelihood-based analysis of repeated categorical responses

Jukka Jokinen *,[1]

Department of Vaccines, National Public Health Institute, Helsinki, Finland; Department of Mathematics and Statistics, University of Helsinki, Finland

Abstract

Likelihood-based marginal regression modelling for repeated, or otherwise clustered, categorical responses is computationally demanding. This is because the number of measures needed to describe the associations within a cluster increase geometrically with increasing cluster size. The proposed estimation methods typically describe the associations using odds ratios, which result in computationally unfeasible solutions for large cluster sizes. An alternative method for joint modelling of the regression, association, and dropout mechanism for clustered categorical responses is presented. The joint distribution of a multivariate categorical response is described by utilizing the mean parameterization, which facilitates maximum likelihood estimation in two important respects. The models are illustrated by analyses of the presence and abscence of schizophrenia symptoms on 86 patients at 12 repeated time-points, and a survey of opinions of 607 adults regarding government spending on 9 different targets, measured on a common 3-level ordinal scale. Free software is available.

Key words: Repeated categorical data, Mean parameterization, Dependence ratio, Marginal model, Association model, Selection model

1 Introduction

Marginal regression analysis of repeated, or otherwise clustered, categorical responses typically focus on quantifying the population-averaged effects of

* National Public Health Institute, Mannerheimintie 166, 00300 Helsinki, Finland
Email address: `jukka.jokinen@helsinki.fi` (Jukka Jokinen).
URL: `http://www.helsinki.fi/~jtjokine/drm` (Jukka Jokinen).
[1] I thank Anders Ekholm for helpful comments and the Southampton Statistical Sciences Research Institute (UK) for providing computing facilities.

the explanatory variables on the responses. Two examples of this type are presented. In the first example, the response of interest is the presence or abscence of schizophrenia symptoms of 86 patients at 12 repeated time-points. The focus is to find out whether the longitudinal symptom prevalence differs according to age-at-onset and sex. The second example is a survey of 607 subjects, regarding opinions about government spending on 9 different targets. Each of the responses is ordinal, with levels 0=too little, 1=about right, 2=too much. Here the focus is to examine whether the opinions of subjects differ according to their political party affiliation. Standard methods, relying on assumptions of independence, are no longer valid for the analyses of these datasets, since repeated measurements within a cluster are unlikely to be independent of each other.

Consider the clustered responses in the government spending example: denote the response Y_{ij}, for individual $i = 1, \ldots 607$, and government spending target $j = 1, \ldots, 9$. One possible solution to obtain independent observations is to treat the individual's whole response profile, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i9})$ as the response variable. This requires the use of multivariate probability distributions for the responses. Compared to multivariate normal responses, the association structure of multivariate categorical responses is much more complex. For example, the probability distribution of a multivariate normal response profile of length 9 is fully specified with 9 means, 9 variances and $\{9 \times (9-1)\}/2 = 36$ correlation coefficients. However, for the government spending example, $9(3-1) = 18$ means and $3^9 - 9(3-1) - 1 = 19\,664$ association measures are required. Because of this complexity, likelihood-based marginal regression analysis is generally unfeasible for such large cluster sizes, especially with several explanatory variables.

Proposed methods for likelihood-based analysis typically describe the associations using odds ratios. For a multivariate binary response, Fitzmaurice and Laird (1993) proposed the so-called mixed parameterization, where the univariate marginal probabilities are complemented with conditional log odds ratios, which are the canonical parameters of the log-linear model. Heumann (1996) extended the mixed parameterization approach to ordinal and nominal responses, and Kastner et al. (1997) provided software for fitting these models, based on the iterative proportional fitting algorithm (IPF; Deming and Stephan, 1940). McCullagh and Nelder (1989, Sec. 6.5.) parameterized the associations using logistic factorial contrasts, where, for binary responses, the second order association measure is the marginal odds ratio. For ordinal responses, this parameterization coincides with global odds ratios for the second order associations as proposed by Dale (1986). Glonek and McCullagh (1995) developed a Newton-type iteration algorithm for fitting such models, whereas Dale (1986) and Molenberghs and Lesaffre (1994) utilized the distribution proposed by Plackett (1965) for specifying the joint distribution. Glonek (1996) proposed a hybrid of marginal and conditional odds ratio parameterizations as a computationally more feasible substitute for the approach in Glonek and McCullagh (1995). Building on methods for multicategorical responses, Lang and Agresti (1994) described the associations using local odds ratios and proposed a fitting algorithm using restricted Lagrangian likelihood equations combined with the Newton-Raphson method. Lang (2004) considered a much broader class of models, encompassing the method presented in Lang and Agresti (1994), and also provided software for estimating such models.

In this paper, an alternative to odds ratios is presented for specifying the joint distribution. This method leads to a computationally more efficient algorithm, enabling maximum likelihood (ML) estimation of such large cluster sizes as in the government spending dataset. For a multivariate binary response, Ekholm et al. (1995) utilized the mean parameterization and specified the joint probability with univariate probabilities and dependence ratios. The dependence ratio is defined as the joint success probability divided by the joint success probability assuming independence. For example, the second order dependence ratio is

$$\tau_{12} = \frac{\mathrm{pr}(Y_1 = 1, Y_2 = 1)}{\mathrm{pr}(Y_1 = 1)\,\mathrm{pr}(Y_2 = 1)}. \tag{1}$$

In words, τ_{12} measures how many times more probable it is to observe both $\{Y_1 = 1\}$ and $\{Y_2 = 1\}$ than would be expected if the two events were independent. This straighforward interpretation also generalises to dependence ratios of all orders. Note also that in contrast to the more commonly used odds ratio, the dependence ratio is an event-specific association measure: rather than measuring the dependence between two random variates Y_1 and Y_2, τ_{12} measures the dependence between the events $\{Y_1 = 1\}$ and $\{Y_2 = 1\}$. An extensive discussion regarding this and other properties of the dependence ratio can be found in Ekholm et al. (2000, 2003), and, particularly, a comparison with the odds ratio in Ekholm (2003).

The dependence ratio parameterization for multivariate binary, nominal, or ordinal responses is presented in Section 2, where also the explicit connection between the model parameters and the joint probability of the response profile is derived. This results in fewer iterative steps in the estimation procedure than for models based on odds ratio parameterisations. Moreover, unrestricted maximisation of the likelihood can be used, since the mean parameterization has the inherent property that the cell probabilities sum to unity. This feature, facilitating the estimation further, is discussed in Section 3. The applicability of the method is illustrated by analysing the two example datasets in Sections 4 and 5, and finally, Section 6 holds a comparison with other proposed methods for marginal modelling of clustered categorical responses.

2 The mean parameterization and the dependence ratio

Denote the response of unit $i = 1, \ldots, n$ at subunit $j = 1, \ldots, q$ by Y_{ij}, with f possible categorical realisations $a_j = 0, \ldots, f-1$. The scores for the categories can be purely nominal, but the dependence ratio parameterisation applies equally for responses measured on an ordinal scale. Note that with $f = 2$, the response is binary. In order to quantify population-averaged effects of the explanatory variables, the regression model parameters need to be related to univariate pointwise or cumulative probabilities, respectively, $\text{pr}(Y_{ij} = a)$ or $\text{pr}(Y_{ij} \leq a)$. Denote a suitable link function, e.g. logit, probit or complementary log-log, of $\text{pr}(Y_{ij} = a)$ or $\text{pr}(Y_{ij} \leq a)$, by $\eta_{ij}(a)$. The marginal regression model is then of the form

$$\eta_{ij}(a) = \theta_a + \boldsymbol{\beta}_a \boldsymbol{x}_{ij}^T, \tag{2}$$

for $a = 0, \ldots, f-2$, where θ_a are the intercept parameters, $\boldsymbol{\beta}_a$ vectors of regression coefficients, constant with respect to i and j, and $\boldsymbol{x}_{ij}$ a vector of explanatory values. For cumulative probabilities, the regression parameters are often assumed to be constant over the response categories, i.e. $\boldsymbol{\beta}_a = \boldsymbol{\beta}$.

The univariate probabilities are, however, not sufficient for specifying the joint probability of a multivariate categorical response. In order to make ML estimation possible, probabilities of f^q alternative realisations of a profile are required. Denote the probability of a particular response profile by $\pi_i(a_1, \ldots, a_q) = \text{pr}(Y_{i1} = a_1, \ldots, Y_{iq} = a_q)$, for $a_j = 0, \ldots, f-1$, with the f^q probabilities of unit i summing to 1. By assuming an underlying multinomial distribution, the log-likelihood function can now be written as

$$\sum_{i=1}^{n} \log\{\pi_i(a_{i1}, \ldots, a_{iq})\}. \tag{3}$$

The probabilities in (3) are sometimes called multinomial cell probabilities, but in the current context, they are referred to as profile probabilities. Note that the marginal regression model parameters in (2) refer to univariate marginal or cumulative probabilities, but the log-likelihood function (3) refers to multivariate profile probabilities. In a survey of clustered categorical data analysis, Agresti and Natarajan (2001) stated that 'hence, one cannot directly substitute the model formula in the log-likelihood function and maximize using standard methods'. This generally applies to methods based on odds ratio parameterizations. In what follows, a general explicit formula is derived for computing $\pi_i(a_1, \ldots, a_q)$ with given marginal regression and association parameters, enabling maximisation of the log-likelihood using standard methods.

2.1 The mean parameterization

Ekholm et al. (1995) derived a one-to-one correspondence between the mean parameters and the profile probabilities for multivariate binary responses. For multivariate nominal responses, Teugels and Van Horebeek (1998) considered the mean parameterization as an analysis tool, complementary to log-linear models. While also useful as an alternative to the log-linear approach, the mean parameterization is especially handy for marginal modelling of multivariate categorical responses, since the correspondence between the univariate probabilities and the mean parameters is straightforward. To demonstrate this connection concurrently for binary, nominal and ordinal responses, define $f-1$ dummy variables, for $a = 1, \ldots, f-1$, by $Y_i^{(a)} = 1$ if $Y_i = a$, else 0. Denote the moments of the responses, generally, by μ. The univariate means are then,

$$\begin{aligned} \mathrm{pr}(Y_i = a) &= E(Y_i^a) = \mu_i^{(a)}, \text{ for } a = 1, \ldots, f-1, \\ \mathrm{pr}(Y_i = 0) &= 1 - \mu_i^{(1)} - \cdots - \mu_i^{(f-1)}. \end{aligned} \tag{4}$$

For a general correspondece between moments and profile probabilities, it is helpful to express the equations (4) in matrix form. Define the $f \times f$ matrix W as

$$W = \begin{pmatrix} 1 & \text{-1} & \ldots & \text{-1} \\ 0 & 1 & \ldots & 0 \\ & & \vdots & \\ 0 & 0 & \ldots & 1 \end{pmatrix}. \tag{5}$$

The connection in (4) can thus be expressed as

$$\{\mathrm{pr}(Y_i = 0), \ldots, \mathrm{pr}(Y_i = f-1)\} = (1, \mu_i^{(1)}, \ldots, \mu_i^{(f-1)})\, W^T. \tag{6}$$

For a general q-variate case, second and higher order means are defined as

$$\begin{aligned} \mu_{ijk}^{(a_j,a_k)} &= E(Y_{ij}^{(a_j)} Y_{ik}^{(a_k)}) = \mathrm{pr}(Y_{ij} = a_j, Y_{ik} = a_k), \\ &\vdots \\ \mu_{i1\ldots q}^{(a_1,\ldots,a_q)} &= E(Y_{i1}^{(a_1)} \cdots Y_{iq}^{(a_q)}) = \mathrm{pr}(Y_{i1} = a_1, \ldots, Y_{iq} = a_q), \end{aligned}$$

for $a_j, a_k = 1, \ldots, f-1$ and $j, k = 1, \ldots, q$ with $j < k$. The complete list of means, or moments of order $1, \ldots, q$, can be conveniently expressed with the

Kronecker product, $\otimes$. Define the $1 \times f^q$ vector $\boldsymbol{\mu}_i$ as

$$\boldsymbol{\mu}_i = E\{(1, Y_{iq}^{(1)}, \ldots, Y_{iq}^{(f-1)}) \otimes \cdots \otimes (1, Y_{i1}^{(1)}, \ldots, Y_{i1}^{(f-1)})\},$$

and the corresponding $1 \times f^q$ vector of profile probabilities as

$$\boldsymbol{\pi}_i = \{\pi_i(0, 0, \ldots, 0), \pi_i(1, 0, \ldots, 0), \ldots, \pi_i(f-1, f-1, \ldots, f-1)\},$$

where the first response is varying fastest and the last response slowest. Further denote the Kronecker product of q identical matrices (5) by $W_q = W \otimes \cdots \otimes W$. It follows that the q-variate generalisation of (6) is obtained as

$$\boldsymbol{\pi}_i = \boldsymbol{\mu}_i W_q^T. \tag{7}$$

For example, if $f = 2$ and $q = 3$:

$$\begin{pmatrix} \pi_i(0,0,0) \\ \pi_i(1,0,0) \\ \pi_i(0,1,0) \\ \pi_i(1,1,0) \\ \pi_i(0,0,1) \\ \pi_i(1,0,1) \\ \pi_i(0,1,1) \\ \pi_i(1,1,1) \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \mu_{i1}^{(1)} \\ \mu_{i2}^{(1)} \\ \mu_{i12}^{(1,1)} \\ \mu_{i3}^{(1)} \\ \mu_{i13}^{(1,1)} \\ \mu_{i23}^{(1,1)} \\ \mu_{i123}^{(1,1,1)} \end{pmatrix}.$$

Note that for $f = 2$, that is, the binary case, the superscripts of μ_i do not vary, and are therefore dropped in subsequent notations.

2.2 The dependence ratio

To complement the univariate means with more interpretable measures of the associations, Ekholm et al. (1995, 2003) replaced the second- and higher order means with dependence ratios, which are defined as

$$\tau_{jk}^{(a_j,a_k)} = \frac{\mu_{ijk}^{(a_j,a_k)}}{\mu_{ij}^{(a_j)}\mu_{ik}^{(a_k)}},$$

$$
\begin{aligned}
\tau_{jkl}^{(a_j,a_k,a_l)} &= \frac{\mu_{ijkl}^{(a_j,a_k,a_l)}}{\mu_{ij}^{(a_j)}\mu_{ik}^{(a_k)}\mu_{il}^{(a_l)}},\\
&\vdots\\
\tau_{1\ldots q}^{(a_1,\ldots,a_q)} &= \frac{\mu_{i1\ldots q}^{(a_1,\ldots,a_q)}}{\mu_{i1}^{(a_1)}\cdots\mu_{iq}^{(a_q)}},
\end{aligned}
\tag{8}
$$

for $a_j, a_k, a_l = 1, \ldots, f-1$ and $j, k, l = 1, \ldots, q$ with $j < k < l$, that is, the joint probability divided by the joint probability assuming independence. It is apparent from (8) that the second and higher order means are simple transformations of the univariate means and dependence ratios. Denote the transformed $1 \times f^q$ vector $\boldsymbol{\mu}_i$, consisting of univariate means and dependence ratios of all orders, by $\boldsymbol{\psi}_i = (\boldsymbol{\mu}_{1i}, \boldsymbol{\tau})$. For example, if $f = 2$ and $q = 3$,

$$
\boldsymbol{\psi}_i = (1,\ \mu_{i1},\ \mu_{i2},\ \mu_{i1}\mu_{i2}\tau_{12},\ \mu_{i3},\ \mu_{i1}\mu_{i3}\tau_{13},\ \mu_{i2}\mu_{i3}\tau_{23},\ \mu_{i1}\mu_{i2}\mu_{i3}\tau_{123}).
$$

It follows from (8) and (7) that

$$
\boldsymbol{\pi}_i = \boldsymbol{\psi}_i W_q^T. \tag{9}
$$

2.3 Association structures

Vector $\boldsymbol{\psi}_i$ is composed by $q(f-1)$ univariate means and $f^q - q(f-1) - 1$ dependence ratios. To achieve parsimonious modelling of the associations, a structure needs to be imposed on the dependence ratios. The following equations demonstrate how temporal and exchangeable association structures can be formulated using dependence ratios. For details and discussion of these structures, see Ekholm et al. (2000, 2003); Jokinen et al. (2006). Denote the vector of association model parameters by $\boldsymbol{\alpha}$. In order to simplify the expression of τ for exchangeable structures, let $|w_a|$ denote the number of repetitions of the response categories $a = 1, \ldots, f-1$. For example, for $\tau^{(2,1,4,1,4,4)}$, $w_1 = 2, w_2 = 1, w_3 = 0$ and $w_4 = 3$, and thus $\tau^{(|w_1|,|w_2|,|w_3|,|w_4|)} = \tau^{(|2|,|1|,|0|,|3|)}$. Further denote by $|w| = w_1 + \cdots + w_{f-1}$, for $w = 2, \ldots, q$.

($\mathcal{N}$): Suppose that each unit i does or does not carry a factor necessary for $a > 0$. Furthermore, suppose that the subunits are conditionally independent given the necessary factor. Denote the probability that the necessary factor is present by $\mathrm{pr}(N_i = 1) = \nu_1$. The τ's of all orders simplify to

$$
\tau^{(a_1,\ldots,a_w)} = \tau^{|w|} = 1/\nu_1^{w-1}.
$$

This association model has a single parameter $\alpha = \nu_1$.

($\mathcal{L}$): Suppose that each unit i has a realisation of a latent factor, $L_i = 0$ or 1, and that subunits are conditionally independent given L. Denote by $\nu_2 = \text{pr}(L_i = 1)$, and by $\kappa^{(a)} = \text{pr}(Y_{ij} = a|L_i = 0)/\text{pr}(Y_{ij} = a|L_i = 1)$, for $j = 1, \ldots, q$. The τ's have the form

$$\tau^{(a_1,\ldots,a_w)} = \frac{\nu_2 + (1-\nu_2)\kappa^{(a_1)} \cdots \kappa^{(a_w)}}{\{\nu_2 + (1-\nu_2)\kappa^{(a_1)}\} \cdots \{\nu_2 + (1-\nu_2)\kappa^{(a_w)}\}},$$

for $a = 1, \ldots, f-1$. This model has f parameters $\boldsymbol{\alpha} = (\nu_2, \kappa^{(1)}, \ldots, \kappa^{(f-1)})$. Note that if $\kappa^{(1)} = \cdots = \kappa^{(f-1)} = 0$, $\mathcal{L}$ simplifies to $\mathcal{N}$.

($\mathcal{D};\mathcal{B}$): Suppose that each unit i has a realisation of a vector of latent continuous variables $\boldsymbol{P}_i = (P_{i1}, \ldots, P_{i(f-1)})$, and that subunits are conditionally independent given $\boldsymbol{P}_i$. Assume further that $\boldsymbol{P}_i \sim \textit{Dirichlet}(\xi_0, \ldots, \xi_{f-1})$. It follows that, for all $w_a > 0$,

$$\tau^{(|w_1|,\ldots,|w_{f-1}|)} = \frac{\prod_{k=1}^{w_1} \xi_1 + k - 1}{\xi_1^{w_1}} \times \cdots \times \frac{\prod_{k=1}^{w_{f-1}} \xi_{f-1} + k - 1}{\xi_{f-1}^{w_{f-1}}} \times \frac{(\xi_0 + \cdots + \xi_{f-1})^w}{\prod_{k=1}^{w} \xi_0 + \cdots + \xi_{f-1} + k - 1},$$

with f parameters $\boldsymbol{\alpha} = (\xi_0, \ldots, \xi_{f-1})$. Note that if $f = 2$, then P_i is a scalar, and the Dirichlet distribution simplifies to Beta distribution: $P_i \sim \textit{Beta}(\xi_1, \xi_0)$. Furthermore, τ simplifies to

$$\tau^{|w|} = \frac{\xi_1 + 1}{\xi_0 + \xi_1 + 1} \times \cdots \times \frac{\xi_1 + w - 1}{\xi_0 + \xi_1 + w - 1} \times \left(\frac{\xi_0 + \xi_1}{\xi_1}\right)^{w-1}.$$

($\mathcal{M}$): Suppose that, for longitudinal studies, the responses satisfy the first order Markov assumption: $\text{pr}(Y_{i(j+1)} = a|Y_{i1}, \ldots, Y_{ij}) = \text{pr}(Y_{i(j+1)} = a|Y_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, q-1$. It follows that the profile probabilities can be expressed through univariate means and overlapping bivariate probabilities:

$$\pi_i(a_1, \ldots, a_q) = \frac{\text{pr}(Y_{i1} = a_1, Y_{i2} = a_2) \cdots \text{pr}(Y_{i(q-1)} = a_{q-1}, Y_{iq} = a_q)}{\text{pr}(Y_{i2} = a_2) \cdots \text{pr}(Y_{i(q-1)} = a_{q-1})}, \quad (10)$$

for $a = 0, \ldots, f-1$. Equation (10) implies that $(f-1)^2(q-1)$ second order dependence ratios $\boldsymbol{\alpha} = (\tau_{12}^{(1,1)}, \ldots, \tau_{(q-1)q}^{(f-1,f-1)})$ are needed for this assocation model.

($\mathcal{M}2$): Extending $\mathcal{M}$ for longitudinal studies, assume the second order Markov property: $\text{pr}(Y_{i(j+1)} = a|Y_{i1}, \ldots, Y_{ij}) = \text{pr}(Y_{i(j+1)} = a|Y_{ij}, Y_{i(j-1)})$. The profile probabilities can now be expressed through overlapping bi- and trivariate probabilities. For each trivariate probability, $f^3 - 3(f-1) - 1$ τ's are required. Already for a 3-level response, this amounts to 20 parameters. Therefore, arguably, $\mathcal{M}2$ is only feasible for a binary response, whence $3q - 5$ parameters $\boldsymbol{\alpha} = (\tau_{12}, \tau_{13}, \tau_{23}, \tau_{123}, \ldots, \tau_{(q-2)(q-1)q})$ are required.

These association structures can be extended by introducing functional forms for the variation over time of the dependence ratios (Ekholm et al., 2002; Jokinen et al., 2006), as well as by including explanatory variables. In addition, combinations of the association models can be imposed. For a vector of covariates $\boldsymbol{z}_i$, and time indicators $\boldsymbol{t} = (t_1, \ldots, t_q)$, denote the set of equations for $\{\mathcal{N}, \mathcal{L}, \mathcal{B}, \mathcal{D}, \mathcal{M}, \mathcal{M}2, \mathcal{NL}, \mathcal{NB}, \mathcal{ND}, \mathcal{NM}, \mathcal{LM}, \mathcal{NLM}, \mathcal{NM}2\}$ by

$$\boldsymbol{\tau}_i = g(\boldsymbol{\alpha}; \boldsymbol{z}_i, \boldsymbol{t}). \tag{11}$$

Note from (2), (11) and (9) that a general explicit solution exists for the profile probabilities in terms of the joint regression and association parameters: $\boldsymbol{\psi}_i = \{\boldsymbol{\mu}_1(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{x}_{ij}), \boldsymbol{\tau}(\boldsymbol{\alpha}; \boldsymbol{z}_i, \boldsymbol{t})\}$.

2.4 Drop-out in longitudinal studies

In longitudinal studies, it is common that some units drop out before completing the study. In that case, the influence of the drop-out can be examined by adding a selection model (Diggle and Kenward, 1994) on top of the joint regression and association model. The conditional hazard for dropping out at measurement point $j = 2, \ldots, q$ is defined as $\mathrm{pr}(D_i = j | D_i \geq j) = \phi(\boldsymbol{\delta}; j, \boldsymbol{a}_i, \boldsymbol{v}_i)$. The effect of the covariates are regressed to the drop-out hazard as follows:

$$\mathrm{logit}\{\phi(\boldsymbol{\delta}; j, \boldsymbol{a}_i, \boldsymbol{v}_i)\} = \delta_0 + \delta_1 a_{ij} + \delta_2 a_{i(j-1)} + \boldsymbol{\delta}_3 \boldsymbol{v}_{i(j-1)}, \tag{12}$$

where $\boldsymbol{v}_{i(j-1)}$ is a vector of explanatory variables at $j - 1$. Three types of missing data mechanisms (Little and Rubin, 2002), that is, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), can be investigated using parameter restrictions on (12). Imposing a restriction $\delta_1 = 0$ implies MAR mechanism, whereas $\delta_1 = \delta_2 = \boldsymbol{\delta}_3 = 0$ implies MCAR. This drop-out model can be specified on top of the joint regression and association model, resulting in an explicit formula for the likelihood in terms of all three parts of the model: $\boldsymbol{\psi}_i^* = \{\boldsymbol{\mu}_1(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{x}_{ij}), \boldsymbol{\tau}(\boldsymbol{\alpha}; \boldsymbol{z}_i, \boldsymbol{t}), \phi(\boldsymbol{\delta}; j, \boldsymbol{a}_i, \boldsymbol{v}_i)\}$. See Ekholm et al. (2003, p.798 and p.802) for more details on the theory and an empirical example.

3 Maximum likelihood estimation

As noted by Agresti and Natarajan (2001), methods for maximum likelihood (ML) fitting of marginal and cluster-specific models are difficult to implement or require approximate methods. However, the dependence ratio parameterization provides two improvements that facilitate the implementation of ML estimation: (i) closed-form expression for all the profile probabilities, and (ii) the inherent property that they sum to unity.

3.1 Closed-form expression

In order to obtain ML estimates of the model parameters, successive iterations of the log-likelihood with respect to $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ are required:

$$\boldsymbol{l}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{x}_{ij}, \boldsymbol{z}_i, \boldsymbol{t}) = \sum_{i=1}^{n} \log\{\pi_i(a_{i1}, \ldots, a_{iq})\}. \tag{13}$$

For the odds ratio parameterizations (Fitzmaurice and Laird, 1993; Glonek and McCullagh, 1995; Lang and Agresti, 1994), no explicit solution for computing $\pi_i(a_{i1}, \ldots, a_{iq})$ in terms of $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is generally available. Therefore, the iterative procedure is required to have two alternating steps: 1) finding the next round of values for $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$, and 2) finding the next round of values for $\pi_i(a_{i1}, \ldots, a_{iq})$. In order to implement this two-step procedure, methods like IPF (Deming and Stephan, 1940) and Lagrangian constraints (Haber, 1985), combined with the Newton-Raphson method, have been utilized when using odds ratio parameterisations. For the dependence ratio parameterisation, log-likelihood function (13) can be explicitly calculated for given $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$, thereby requiring only a single step at each round of iteration.

3.2 Inherent unit-sum constraint

Glonek (1996) noted that full-likelihood approaches are generally not suitable for data consisting of large clusters. This is because it is necessary to calculate and manipulate all probabilities of the multinomial distribution, which rapidly becomes prohibitive. One reason for this necessity is to ensure that all f^q profile probabilities for each unit i sum up to one. However, when using the mean parameters, this is of no concern, since the parameterization has an inherent unit-sum constraint. This is most apparent in the univariate binary case, where $\mathrm{pr}(Y_{ij} = 0) = 1 - \mu_{ij}$ constrains the sum of failure and success probabilities to one; see also equation (4). More generally, by multiplying with $1 \times f^q$ vector of ones, denoted by $\mathbf{1}$, equation (7) gives

$$\boldsymbol{\pi}_i \mathbf{1}^T = \boldsymbol{\mu}_i W_q^T \mathbf{1}^T = [1, \mu_{i1}^{(1)}, \ldots, \mu_{i1\ldots q}^{(f-1,\cdots,f-1)}] \times [1, 0, \ldots, 0]^T = 1. \qquad (14)$$

This property leads to an overwhelming improvement in terms of estimation, since only the observed profile probabilities need to be calculated for given parameters at each iteration. For example, consider the government spending dataset, with a unique covariate pattern for each individual: instead of calculating $607 \times 3^9 = 11\,947\,581$ profile probabilities at each iteration, using mean parameters, one only needs to calculate the probabilities of 607 observed profiles. Once the complexity increases, it is apparent how ML fitting using the mean parameterization introduces a vast stride compared to estimation methods that lack this inherent constraint. As a consequence, an unconstrained numerical optimiser can be applied to obtain ML estimates, by utilizing the explicit equation (9). As Lindsey and Lindsey (2006) pointed out, with modern computers and statistical software, nonlinear functions can be fitted as easily as linear ones.

This estimation technique constrains the observed profile probabilities to be positive, and the sum of all possible profile probabilities to one. However, it does not guarantee that the probabilities of all the unobserved profiles are positive. When modelling datasets with complex structures and extensive number of unobserved profiles, the fitted probabilities for some of the unobserved profiles may turn out negative. Certain parameter restrictions could be imposed to avoid this to happen, but unless the restrictions imply a meaningful model, arguably a more sensible modelling approach is to consider the model specification to be inadequate, and reformulate the model until all possible profiles turn out positive. It is also important to recognize that when fitting complex nonlinear models, especially with latent structures and/or postulated dropout mechanisms, problems such as flat likelihood over ranges of the parameter space or multiple maxima may occur (Seber and Wild, 2003). Although in many cases starting the iteration from the assumption of independence is adequate, it is advisable to monitor the iteration process and to try different starting values if necessary.

4 Madras Longitudinal Schizophrenia Study

Diggle et al. (2002, pp.234-43) analyzed a dataset from Madras Longitudinal Schizophrenia Study, where psychiatric symptoms of 86 patients were investigated over the first year after initial hospitalisation for disease. Schizophrenia symptoms were classified as positive (hallucinations, delusions, thought disorders) and negative (flat affect, apathy, withdrawal). Presence or abscence of these symptoms were recorded monthly after hospitalisation. From among the symptoms, Diggle et al. (2002) selected 'thought disorders' (0=no, 1=yes) as the response variable for the analysis. Monthly frequencies of thought disorders of the 86 patients are summarised in Table 1.

Table 1
Monthly frequencies of thought disorders in the Madras Study

Month	0	1	2	3	4	5	6	7	8	9	10	11
No	30	34	40	43	50	56	63	65	65	64	64	63
Yes	56	51	43	37	28	20	13	10	7	8	6	6
Missing	0	1	3	6	8	10	10	11	14	14	16	17

For the re-analysis of this dataset, we follow the regression model formulation of Diggle et al. (2002). The focus is to find out whether the longitudinal symptom prevalence of thought disorders differs across patient subgroups, defined by age (0, if <20 years old, otherwise 1) and sex (0=male, 1=female). Therefore the main interest in the regression model focuses on the interaction between the explanatory variables and month $t_j = 0, \ldots, 11$. The regression model is of form

$$\text{logit}\{\mu_{it_j}\} = \theta_0 + \beta_1\, t_j + \beta_2\, \text{age}_i + \beta_3\, \text{sex}_i + \beta_4\, t_j.\text{age}_i + \beta_5\, t_j.\text{sex}_i. \qquad (15)$$

For the analysis of the associations, temporal structures are the most plausible choices for this type of longitudinal response. For an exploratory investigation of the length of the stochastic memory, we fitted three standard logistic regression models: in addition to the covariates in (15), we included previous responses (i) $(y_{it_{j-1}})$, (ii) $(y_{it_{j-1}}, y_{it_{j-2}})$, and (iii) $(y_{it_{j-1}}, y_{it_{j-2}}, y_{it_{j-3}})$, as explanatory variables. Akaike's Information Criteria (AIC) for the models are, respectively, 726.5, 727.0 and 729.0, suggesting that the first order Markov assumption is adequate for the associations. Therefore, for the combined modelling task, we specified the regression as in (15) and the association initially assuming $\mathcal{M}$, the model denoted by $\{R, \mathcal{M}\}$.

The observed second order dependence ratios $\tau_{t_j t_{j+1}}$, $t_j = 0, \ldots, 10$, are plotted in Figure 1. The dependence ratios increase rapidly by pairs of months, suggesting that the independence assumption is clearly inadequate. For example, the observed probability of having thought disorders at both months ten and eleven is 11.5 times greater than would be expected if the two events were independent. An alternative interpretation is in terms of conditional probabilities: $\tau_{t_j t_{j+1}} = \text{pr}(Y_{i\,t_{j+1}} = 1 | Y_{i\,t_j} = 1)/\text{pr}(Y_{i\,t_{j+1}} = 1)$, i.e., the probability of thought disorders at month eleven given the disorders were present at month ten, is 11.5 times the marginal probability.

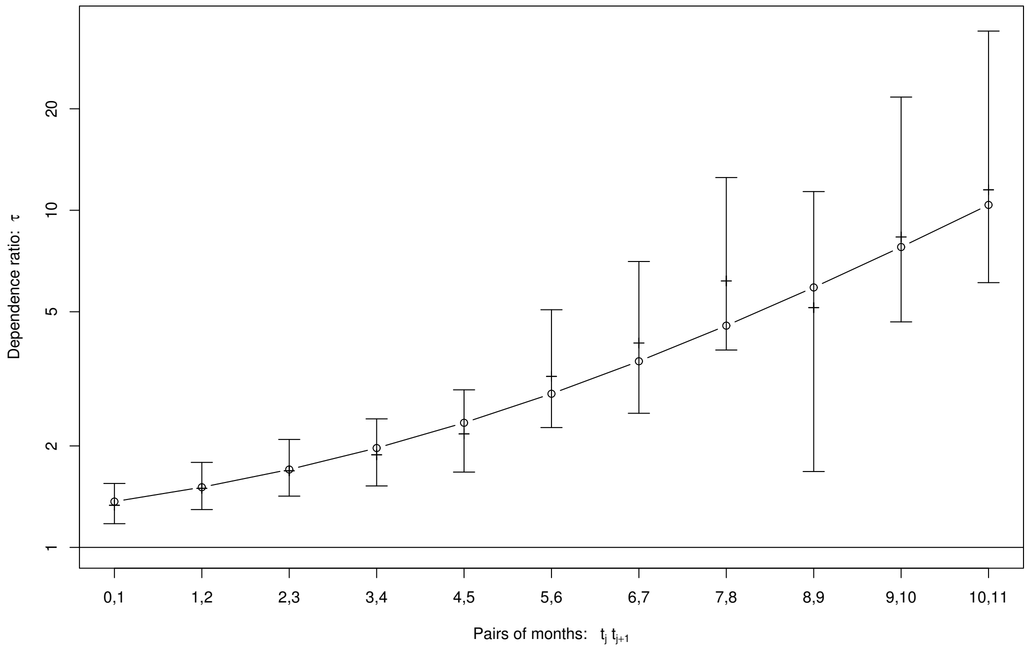

Fig. 1. Dependence ratios $\tau_{t_j t_{j+1}}$ in the Madras Schizophrenia Study: Observed values with bootstrap 95% CI's (error bars) compared with the fitted values (−o−) from the model $\{R, \mathcal{M}_2\}$. Horizontal line corresponds to independence.

Since estimating 10 dependence ratios from the data lacks parsimony and interpretation, we explored functional forms for the adjacent dependence ratios. The relation $\tau_{t_j t_{j+1}} = 1 + \alpha_0 \exp(\alpha_1 t_j)$ for the second order τ's fits the observed dependence ratios almost perfectly, as can be seen from Figure 1. This model $\{R, \mathcal{M}_2\}$, where the subscript on $\mathcal{M}$ indicates the number of additional parameters compared to the model assuming independence $\{R, \mathcal{I}\}$, produces an overwhelming improvement in the fit (AIC's respectively 692.5 and 929.3).

There were altogether 17/86 subjects that dropped out from the study before the end of the follow-up at month 11. We investigated the effect of the drop-out on the regression and association estimates by fitting a model for the non-response mechanism as specified in equation (12). The estimates of the drop-out model did not support the MNAR assumption, although it has to be noted that the effect of the drop-out can hardly be realistically investigated on the basis of only 17 drop-outs.

Six different types of symptoms were investigated in the Madras study. Therefore, it is conceivable that there might be variation in the positive and negative symptom profiles of the hospitalised patients. For example, there may be a certain subgroup that are hospitalised because of e.g. negative symptoms, and thus are not susceptible to such positive symptoms as thought disorders during the first year after hospitalisation. Furthermore, the size of this subgroup may vary across patient subgroups. This type of association generating mechanism can be formulated by superimposing $\mathcal{N}$ on top of $\mathcal{M}$, resulting in a joint model $\{R, \mathcal{N}\mathcal{M}\}$. The τ's for adjacent months have then the form $\tau_{t_j t_{j+1}} = \{1 + \alpha_0 \exp(\alpha_1 t_j)\}/\mathrm{pr}(N_i = 1)$; see Section 2.3. Furthermore, by allowing the proportions of the susceptibles to vary across patient subgroups, the second order dependence ratios can now be specified as

$$\tau_{i\,t_j\,t_{j+1}} = \frac{1 + \alpha_0 \exp\{\alpha_1 t_j\}}{\alpha_2 + \alpha_3\,\mathrm{sex}_i + \alpha_4\,\mathrm{age}_i + \alpha_5\,\mathrm{sex.age}_i}. \tag{16}$$

The joint model $\{R, \mathcal{N}\mathcal{M}_6\}$, specified by (15) and (16) together, fits the data clearly better than $\{R, \mathcal{M}_2\}$. According to AIC (=685.2), the model also fits better than the ones fitted to the same dataset in Diggle et al. (2002). The parameter estimates are summarised in Table 2.

Table 2
ML estimates of model $\{R, \mathcal{N}\mathcal{M}_6\}$ in the Madras Study

	Parameter	Estimate	Std.Error
Regression model	θ_0	0.789	0.269
	β_1	−0.319	0.045
	β_2	0.449	0.370
	β_3	−0.576	0.359
	β_4	−0.038	0.043
	β_5	0.008	0.070
Association model	α_0	0.179	0.055
	α_1	0.390	0.048
	α_2	0.931	0.062
	α_3	−0.251	0.102
	α_4	0.021	0.085
	α_5	0.115	0.081

We conclude from the regression model that there are no indications of interaction between month and age ($\hat{\beta}_4 = 0.038$, s.e. 0.043) or month and sex ($\hat{\beta}_5 = 0.008$, s.e. 0.070). However, when examining the longitudinal prevalence profiles across patient subgroups, not only the marginal univariate probabilities in time, but also the probabilities of the whole response profiles may differ. According to the association model, there is little indication that there is a subgroup of males that are not susceptible to thought disorders ($\hat{\alpha_2} = 0.931$, s.e. 0.062). However, among females, especially at the age less than 20 years, only $100\% \times \{\hat{\alpha_2} + \hat{\alpha_3}\} = 68\%$ are susceptible to thought disorders during the first year after hospitalisation. For females older than 20 years, the susceptibility is getting more prevalent ($100\% \times \{\hat{\alpha_2} + \hat{\alpha_3} + \hat{\alpha_4} + \hat{\alpha_5}\} = 84\%$). In addition, the interpretation of the temporal association is that, for the susceptibles, the history of thought disorders affecting the current state only reaches back one month, with dependence ratios rapidly increasing by increasing pairs of months; see Figure 1.

Note also from Figure 1 and equation (1) that, as an event-specific association measure, the dependence ratio is measuring the association of the presence of thought disorders, rather than abscence of them. Since six different psychiatric symptoms were investigated in the study, zero here can be seen as a collapsed category of different psychiatric states, that is, abscence of all symptoms or presence of other symptoms than thought disorders. Therefore, when the substantial emphasis is on patients suffering from thought disorders, the dependence ratio is focused on the association of interest.

5 General Social Survey: opinions regarding government spending

The dataset comes from the 1989 US General Social Survey, where 607 adults, aged over 18 years, were asked their opinions regarding government spending on 1: National Defence, 2: Assistant to Big Cities, 3: Law Enforcement, 4: Drug Rehabilitation, 5: Education, 6: Environment, 7: Assistant to Poor, 8: Health, and 9: Space Exploration. This dataset, summarised in Table 3, has been previously analysed in parts by Lang and Agresti (1994) and Jokinen et al. (2006).

Each of the responses is ordinal (0=too little, 1=about right, 2=too much), so we apply the proportional odds model formulation (McCullagh and Nelder, 1989, Sec 5.2.2.) for the regression. The regression analysis task is to compare the univariate cumulative probabilities of the 9 government spending targets by political party affiliation. The original categories for party affiliation were (a) Strong Democrat, (b) Not very strong Democrat, (c) Independent, close to Democrat, (d) Independent, (e) Independent, close to Republican, (f) Not very strong Republican, and (g) Strong Republican. Since maintaining these 7 categories would require altogether $7 \times 9 - 1 = 62$ parameters for estimating the regression effects, for parsimony, some collapsing of classes of this particular explanatory variable was thus necessary. The initial exploration of the data revealed that subjects within the levels of each party affiliation were more similar than between parties and thus this variable was collapsed to a 3-level factor with levels: Democrat (a, b), Independent (c, d, e), and Republican (f, g). In order to adjust for possible confounders, we also included age

Table 3
Marginal frequencies (percentages) of responses regarding government spending by political party affiliation.

Government Spending	Party Affiliation	Too Little	About Right	Too Much	Missing
National Defence	Democrat	35 (15.4)	86 (37.7)	98 (43.0)	9 (3.9)
	Independent	29 (16.7)	65 (37.4)	76 (43.7)	4 (2.3)
	Republican	35 (17.1)	108 (52.7)	59 (28.8)	3 (1.5)
Assistance to Big Cities	Democrat	61 (26.8)	94 (41.2)	73 (32.0)	0 (0.0)
	Independent	38 (21.8)	66 (37.9)	70 (40.2)	0 (0.0)
	Republican	35 (17.1)	80 (39.0)	90 (43.9)	0 (0.0)
Law Enforcement	Democrat	156 (68.4)	58 (25.4)	14 (6.1)	0 (0.0)
	Independent	106 (60.9)	55 (31.6)	13 (7.5)	0 (0.0)
	Republican	116 (56.6)	76 (37.1)	13 (6.3)	0 (0.0)
Drug Rehabilitation	Democrat	154 (67.5)	44 (19.3)	20 (8.8)	10 (4.4)
	Independent	119 (68.4)	36 (20.7)	16 (9.2)	3 (1.7)
	Rebuplican	111 (54.1)	75 (36.6)	17 (8.3)	2 (1.0)
Education	Democrat	175 (76.8)	51 (22.4)	2 (0.9)	0 (0.0)
	Independent	134 (77.0)	33 (19.0)	7 (4.0)	0 (0.0)
	Republican	137 (66.8)	59 (28.8)	9 (4.4)	0 (0.0)
Environment	Democrat	164 (71.9)	49 (21.5)	15 (6.6)	0 (0.0)
	Independent	137 (78.7)	29 (16.7)	8 (4.6)	0 (0.0)
	Republican	143 (69.8)	50 (24.4)	12 (5.9)	0 (0.0)
Assistance to Poor	Democrat	176 (77.2)	37 (16.2)	14 (6.1)	1 (0.4)
	Independent	125 (71.8)	36 (20.7)	10 (5.7)	3 (1.7)
	Republican	101 (49.3)	72 (35.1)	29 (14.1)	3 (1.5)
Health	Democrat	186 (81.6)	32 (14.0)	10 (4.4)	0 (0.0)
	Independent	125 (71.8)	35 (20.1)	14 (8.0)	0 (0.0)
	Republican	123 (60.0)	71 (34.6)	11 (5.4)	0 (0.0)
Space Exploration	Democrat	28 (12.3)	86 (37.7)	107 (46.9)	7 (3.1)
	Independent	35 (20.1)	69 (39.7)	66 (37.9)	4 (2.3)
	Rebuplican	30 (14.6)	123 (60.0)	51 (24.9)	1 (0.5)

(in years), sex (male, female), and race (white, black, other) as explanatory variables into the regression model, denoted by R.

There were 4 missing values for party affiliation, and 2 for age. Strictly speaking, multiple imputation (Little and Rubin, 2002) should be used in order to properly take into account the uncertainty arising from these missing explanatory values. However, in this particular dataset, there were $607 \times 9 = 5463$ values of the response variable, each with a 1×30-vector of explanatory values. Therefore we concluded that the effect of using either single or multiple imputation is negligible for only 6 missing values, and these values were thus imputed using single regression imputation. Missing response values (50 out of 5463) were assumed to be MAR.

Lang and Agresti (1994) analyzed the joint regression and association of government spending targets $\{2, 3, 6, 8\}$ without including any covariates, resulting in $3^4 = 81$ possible realisations of the response profile. Jokinen et al. (2006) added targets $\{1, 5, 7\}$, and the above described explanatory variables. The purpose here is to exemplify the geometrically increasing complexity of a clustered categorical response, as well as the versatility of the presented modelling approach by including a total of 9 spending targets, and covariates for the association model. When using the $\mathcal{M}$-structures, as in the Madras Study example in Section 4, the number of association measures increase only linearly with increasing cluster size. However, for the other structures that are exchangeable, this luxury does not apply: $3^7 - (3-1)7 - 1 = 2\,172$ dependence ratios were required to specify the joint distribution of the response profile in Jokinen et al. (2006), and adding two more responses in the current example results in a total of $3^9 - (3-1)9 - 1 = 19\,664$ association measures. In addition, by allowing the parameters of the association model to vary by political party affiliation, the number of dependence measures is multiplied by three.

We fitted models $\{R, \mathcal{I}\}$, $\{R, \mathcal{N}_1\}$, $\{R, \mathcal{L}_3\}$ and $\{R, \mathcal{D}_3\}$, to the government spending dataset. For $\{R, \mathcal{N}_1\}$, $\hat{\nu_1} = 1.00$, suggesting that there is no subgroup of adults that would always answer 0='too little' to a question about government spending. Since $\{R, \mathcal{D}_3\}$ produced negative fitted profile probabilities, and $\{R, \mathcal{L}_3\}$ fitted the data clearly better than $\{R, \mathcal{I}\}$ (AIC's respectively 9146 and 9246), we selected $\mathcal{L}$ as the most plausible structure for the association between repeated responses. Model $\mathcal{L}$ suggests that the population is divided into two groups with different response category probabilities. Furthermore, we fitted a more elaborate model $\{R, \mathcal{L}_9\}$, where the sizes of the latent groups and their corresponding conditional probabilities were further allowed to vary by party affiliation. This resulted in an improved fit, with AIC$=$ 9141.

The investigation of the effects of confounding variables from model $\{R, \mathcal{L}_9\}$ reveal notable marginal differences across subgroups. The estimate for age is 0.0036 (s.e. 0.0015), suggesting that once people get older, the government's financial involvement is considered less important. Furthermore, compared to males, females feel that the government should spend more (-0.183, s.e. 0.057). Racial differences are also significant: with whites as the baseline, the estimates for blacks and other races are, respectively, -0.235 (s.e. 0.099) and 0.449 (s.e. 0.183). For the investigation of the marginal differences of government spending targets, the parameter estimates of the Democrats are contrasted with spending on National Defence (Table 4). The view of the Democrats is that the government should spend less only on Space Exploration than on National Defence. Furthermore, the estimates for the Independents and the Republicans are contrasted with the estimates for the Democrats. In other words, 0 indicates no difference with the views of the Democrats on that specific target. As can be seen from Table 4, the opinions of the Independents generally lie in between the Democrats and the Rebublicans, who in turn differ the most in opinions concerning Assistance to Poor and Health.

Table 4
ML estimates (standard errors) concerning government spending from model $\{R, \mathcal{L}_9\}$. Independents and Rebublicans are contrasted with Democrats.

	Effects	Democrats	Independents	Rebublicans
R	National Defence		$-0.078(0.184)$	$-0.457(0.175)$
	Assistance to Big Cities	$-0.540(0.174)$	$0.259(0.188)$	$0.515(0.204)$
	Law Enforcement	$-2.450(0.188)$	$0.309(0.208)$	$0.353(0.195)$
	Drug Rehabilitation	$-2.502(0.193)$	$0.063(0.223)$	$0.493(0.200)$
	Education	$-2.909(0.199)$	$0.008(0.238)$	$0.411(0.213)$
	Enviromnent	$-2.617(0.191)$	$-0.381(0.236)$	$-0.006(0.209)$
	Assistance to Poor	$-2.897(0.202)$	$0.247(0.235)$	$1.139(0.208)$
	Health	$-3.167(0.211)$	$0.577(0.240)$	$0.944(0.219)$
	Space Exploration	$0.229(0.167)$	$-0.455(0.192)$	$-0.757(0.179)$
$\mathcal{L}_9$	ν_2	$0.338(0.106)$	$-0.127(0.167)$	$0.365(0.120)$
	κ_1	$0.428(0.036)$	$0.036(0.062)$	$-0.174(0.074)$
	κ_2	$1.812(0.321)$	$0.735(1.096)$	$-0.329(0.370)$

The association model estimates are reported in the lower panel of Table 4, where the parameter estimates for the Independents and the Rebublicans are contrasted with the estimates for the Democrats. The interpretation of the model $\mathcal{L}_9$ is that there is no notable difference between the Democrats and the Independents in the way they are divided into two latent groups, denoted by $L = 0$ and $L = 1$. For the Democrats, the proportion in group $L = 0$ is approximately two thirds $(1 - 0.338)$, and in that group, the probability of answering 'about right' is less than half (0.428) of the probability in group $L = 1$. On the contrary, the probability of answering 'too much' is almost twice as probable (1.812) than in group $L = 1$. However, for the Rebublicans, the size of the latent group $L = 0$ is less than one third $(1 - \{ 0.338 + 0.365 \})$. In summary, as a group, the Democrats and the Rebublicans differ both in their marginal distributions, as well as with regard to their reponse profiles.

6 Discussion

Generalized estimation equation technique (GEE, Zeger et al., 1988) is a popular approach for marginal regression analysis of clustered categorical responses. Diggle et al. (2002)[p. 146] stated the following reasons for resorting to GEE: 'We often do not have sensible, simple models for third- and higher-order moments regardless of which formulation we adopt. Even when a probability model is fully specified, the likelihood can be complicated to evaluate except with small and constant n_i' (n_i referring to number of subunits within a unit i). It was demonstrated in Section 2.3 that a rich set of simple and sensible association models for third- and higher-order moments can be specified using the dependence ratio parameterization. It was also shown in Section 3, and illustrated in Sections 4 and 5, that the likelihood evaluation is straightforward, even for relatively large and non-constant number of subunits.

When analysing clustered categorical responses using likelihood-based methods, odds ratios are typically applied for the parameterization of the associations. Conditional odds ratios (Fitzmaurice and Laird, 1993; Heumann, 1996) are problematic for non-constant number of subunits, since the conditioning changes between units. Furthermore, the interpretation of the conditional odds ratios is cumbersome, and therefore they are treated as nuisance parameters. Local odds ratio parameterization (Lang and Agresti, 1994; Lang, 2004) is practically unfeasible when continuous covariates are present, since the univariate means are complex functions of the regression model parameters (Lapp et al., 1998). A useful feature of these two approaches is that software is available for both of these parameterizations, but it has been repeatedly stated (e.g. Heumann, 1996; Agresti and Natarajan, 2001) that these methods are applicable only when the number of subunits is small. In contrast, the dependence ratio parameterization incorporates regression modelling with a meaningful model for the associations for both constant and non-constant number of subunits. Moreover, regression model specification is straightforward also for continuous covariates. In terms of estimation, an important advantage of the method is that there exists an explicit solution for the profile probabilities, and that only the probabilities of the observed profiles have to be computed for finding the ML estimates.

The fact that ML estimates can produce negative fitted probabilities for the unobserved profiles is easily viewed as a limitation of the presented approach.

However, since the estimation method coerces all the observed probabilities positive, estimated negative probabilities can typically occur only for cluster sizes that are generally unfeasible for methods based on odds ratio parameterizations. In addition, the challenge of formulating sensible models is easily forgotten when analysing clustered categorical responses with geometrically increasing complexity. Quoting Diggle et al. (2002)[p. 208]: 'the sensitivity of likelihood-based methods to the underlying model assumptions needs to be understood so that important forms of model violation can be checked in practice'. For example, consider the government spending example in Section 5, with a unique covariate pattern for each subject. Using the terminology familiar to log-linear modelling, the model is fitted to a contingency table with 607 cell counts of one, and 11 946 974 zero cell counts. It is apparent that, for such a sparse contingency table, the joint regression and association model specification requires a careful formulation. If the unconstrained ML estimates produce negative cell probabilities for some of the zero counts, it can be regarded as an indication that the current model specification is inadequate for the problem at hand.

There has been some interesting debate about the advantages and disadvantages of the odds and the dependence ratio parameterization (e.g. Molenberghs and Verbeke, 2004; Ekholm, 2003). Molenberghs and Verbeke (2004), and references therein, advocate the use of global odds ratio parameterization, using an IPF-type algorithm. However, all examples for multicategorical responses, using global odds parameterization, seem to be truncated to a maximum of three repeated measurements, presenting a nominal extension to Dale (1986). It is important to realize that, by utilizing the explicit formula (9) for all possible multivariate probabilities of order $2, \ldots, q$, any type of association measure, including the odds and the dependence ratio, can be derived as a function of the regression and association parameters. Most importantly, from the point of view of the modeller, at least the following questions need to be addressed when choosing the method for the analysis: (i) is the method feasible for the problem at hand, (ii) is the method versatile enough to allow different plausible formulations of the phenomenon under study, and (iii) can the method be easily implemented. While useful as scientific discourse, other comparisons between the methods can be viewed as subordinate to these issues.

A package for freely available software R (Ihaka and Gentelman, 1996), for fitting the models presented in this paper, is available from the author's website. Regression model specification utilizes the flexible regression modelling framework of R. The set of association models presented in Section 2.3, with a possibility of including explanatory variables and functional forms, is also incorporated. Furthermore, selection model specification, again with a possibility of including explanatory variables, is readily implemented. Admittedly, the current version, written completely using the S programming language, is mainly concentrated on flexibility rather than computational speed. Further gains could be achieved by incorporating some compiled code (e.g. C++ or Fortran). Implementation in other statistical softwares is also an interesting prospect. Nevertheless, the two important improvements in the estimation process presented in this paper provide a useful starting point for a more comprehensive likelihood-based analysis of repeated categorical responses.

References

Agresti, A., Natarajan, R. 2001. Modeling Clustered Ordered Categorical Data: a Survey. Int. Statist. Rev. 69, 345-71.

Dale, J.R. 1986. Global cross-ratio models for bivariate, discrete, ordered responses. Biometrics 42, 909-917.

Deming, W.E., Stephan, F.F. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Statist. 11, 427-444.

Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L., 2002. The analysis of Longitudinal Data, second ed. Oxford University Press, Oxford.

Diggle, P.J., Kenward, M.J. Informative dropout in longitudinal data analysis (with discussion). Applied Statistics 43, 49-94.

Ekholm, A., Smith, P.W.F., McDonald, J.W. 1995. Marginal Regression Analysis of a Multivariate Binary Response. Biometrika 82, 847-854.

Ekholm, A., McDonald, J.W., Smith, P.W.F. 2000. Association Models for a Multivariate Binary Response. Biometrics 56, 712-718.

Ekholm A., Jokinen J., Kilpi T. 2002. Combining Regression and Association Modelling for Longitudinal Data on Bacterial Carriage. Statist. Med. 21, 773-791.

Ekholm, A. 2003. Comparing the odds and the dependence ratios. Pp. 13-25 in Statistics, Econometrics and Society: Essays in Honour of Leif Nordberg. R. Höglund, M. Jäntti, and G. Rosenqvist (eds). Helsinki: Statistics Finland.

Ekholm, A., Jokinen, J, McDonald, J.W., Smith, P.W.F. 2003. Joint Regression and Association Modelling for Longitudinal Ordinal Data. Biometrics 59, 795-803.

Fitzmaurice, G.M., Laird, N.M. 1993. A Likelihood-based Method for Analysing Longitudinal Binary Responses. Biometrika 80, 141-151.

Glonek, G.F.V. 1996. A Class of Regression Models for Multivariate Categorical Responses. Biometrika 83, 15-28.

Glonek, G.F.V., McCullagh, P. 1995. Multivariate Logistic Models. J. Roy. Statist. Soc. B 57, 533-546.

Haber, M. 1985. Maximum Likelihood Methods for Linear and Log-Linear Models in Categorical Data. Comput. Statist. Data Anal. 3, 1-10.

Heumann, C. 1996. Marginal regression modeling of correlated multicategorical response: A likelihood approach. SFB386 - Discussion paper 19. Universität München.

Ihaka, R., Gentleman, R. 1996. R: a Language for Data Analysis and Graphics. J. Comput. Graph. Statist. 5, 299-314.

Jokinen, J., McDonald, J.W., Smith, P.W.F. 2006. Meaningful regression and association models for clustered ordinal data. Accepted for publication. Sociological Methodology 36.

Kastner, C., Fieger, A., Heumann, C. 1997. MAREG and WinMAREG A tool for marginal regression models. Comput. Statist. Data Anal. 24, 237-241.

Lang, J.B., Agresti, A. 1994. Simultaneously Modeling Joint and Marginal Distributions of Multivariate Categorical Responses. J. Amer. Statist. Assoc. 89, 625-632.

Lang, J.B. 2004. Multinomial-Poisson homogeneous models for contingency tables. Ann. Statist. 32, 340-383.

Lapp, K., Molenberghs, G., Lesaffre, E. 1998. Models for the association between ordinal variables. Comput. Statist. Data Anal. 28, 387-411.

Lindsey, J.K., Lindsey, P.J. 2006. Multivariate distributions with correlation matrices for nonlinear repeated measurements. Comput. Statist. Data Anal. 50, 720-732.

Little, R.J.A., Rubin, D.B. 2002. Statistical Analysis with Missing Data, second ed. New York, John Wiley.

McCullagh, P., Nelder, J.A. 1989. Generalized Linear Models, second ed. New York: Chapman Hall.

Molenberghs, G., Lesaffre, E. 1994. Marginal Modelling of Correlated Ordinal Data Using a Multivariate Plackett Distribution. J. Amer. Statist. Assoc. 89, 633-644.

Molenberghs, G., Verbeke, G. 2004. Meaningful statistical model formulations for repeated measures. Statistica Sinica 14, 989-1020.

Plackett, R.L. 1965. A class of bivariate distributions. J. Amer. Statist. Assoc. 60, 516-522.

Seber, G.A.F., Wild, C.J. 2003. Nonlinear regression. New York: John Wiley.

Teugels, J.L., Van Horebeek, J. 1998. Algebraic Descriptions of Nominal Multivariate Discrete Data. J. Multivariate Anal. 67, 203-226.

Zeger, S.L., Liang, K.-Y., Albert, P.S. 1988. Models for longitudinal data: a generalized estimating equation approach. Biometrics 44, 1049-1060.