

MARGINAL MODELLING OF MULTIVARIATE CATEGORICAL DATA

GEERT MOLENBERGHS^{1*} AND EMMANUEL LESAFFRE²

¹*Biostatistics, Limburgs Universitair Centrum, B3590 Diepenbeek, Belgium*

²*Biostatistical Center, Katholieke Universiteit Leuven, B3000 Leuven, Belgium*

SUMMARY

This paper describes likelihood methods of analysis for multivariate categorical data. The joint distribution is specified in terms of marginal mean functions, and pairwise and higher order association measures. For the association, the emphasis is on global odds ratios. The method allows flexible formulation of a broad class of designs, such as repeated measurements, longitudinal studies, interrater agreement and cross-over trials. The proposed model can be used for parameter estimation and hypothesis testing. Simple fitting algorithms are proposed. The method is illustrated using a data example. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

Repeated measures studies, and in particular longitudinal studies, are important tools in epidemiological, clinical and social science research. In such studies, the response at each occasion is often recorded as a categorical variable. Further, a categorical outcome frequently specializes to an ordinal variable (for example, no, mild, moderate, or complete relief of pain). Various types of models have been developed, and a wide variety of estimation techniques are proposed. We will consider likelihood methods for estimating the parameters in marginal models, in which the distribution of the marginal responses is modelled as a function of covariates, and, in longitudinal studies, also of time. These models are opposed to conditional models, such as log-linear and graphical models, where the parameters are interpretable in terms of outcome probabilities for a set of outcomes (usually a single outcome), conditional on the outcomes for the other variables.

Consider the following clinical trial. A group of 498 medical students, between 17 and 29 years of age (median 21 years) are randomized to two treatment groups. Those in the HI group receive hepatitis B vaccination (H), followed by influenza vaccination (I), while the reverse order is applied in the IH group. For each type of vaccination, vaccines from both company A and company B are used. In each treatment period, the vaccines are evaluated with respect to the side-effects they caused. We are interested in the outcomes *headache* and *respiratory problems*. Since both outcomes are measured in each of the two periods, we obtain a four-dimensional response variable. It is of interest to assess the strength of the association between both headache

* Correspondence to: Geert Molenberghs, Biostatistics, Center for Statistics, Limburgs Universitair Centrum, B3590 Diepenbeek, Belgium. E-mail: geert.molenberghs@luc.ac.be

outcomes, between both respiratory outcomes, as well as determining whether both complaints are dependent. In addition, a three-point ordinal variable, level of pain, is recorded for six days in a row during the first period, supplementing the cross-over study with a longitudinal one. The first three days will be evaluated here. In order to analyse these data, we need tools for longitudinal categorical data, as well as tools for more complex designs, such as cross-over trials with several outcomes in each period. Whereas the association between outcomes is often considered a nuisance characteristic in longitudinal studies, it is usually of direct interest in multivariate settings, such as the bivariate cross-over study considered here.

The choice between conditional and marginal models frequently arises when analysing multivariate categorical data. Contributions to this discussion can be found in Neuhaus *et al.*¹ and in Liang *et al.*² A choice between various models should be guided by the scientific question that needs to be answered. If a model is chosen, simple expressions are provided for either conditional or marginal probabilities, while the marginal and conditional probabilities, respectively, are complicated functions of the natural parameters of the model.

In fully marginal models, the parameters characterize the marginal probabilities of a subset of the outcomes (ordinarily of a single outcome), without conditioning on the other outcomes. As a result, the effect of an explanatory variable on a particular outcome can be investigated directly, without having to stratify one outcome variable over the others. Further, the design is reproducible in the following sense. If a set of T outcomes $Y = (Y_1, \dots, Y_T)^T$ satisfies a marginal model, then so does every subvector of Y . Also, in a marginal model, the meaning of the parameters does not depend on the number of outcomes; it also does not depend on the presence of higher order parameters (although the numerical value will change). This means that, for example, the result of logistic regressions for each outcome separately is consistent with the picture obtained from a multivariate marginal analysis with logistic marginal distributions.

An advantage of log-linear models is that the parameter vector describing the dependence is not restricted, in the sense that the joint parameter space is the set theoretical product of one-dimensional parameter spaces. The simplest marginal counterexample is the multivariate probit model,³ where the correlations must represent a positive (semi)definite matrix. Further, fitting marginal models tends to be a non-trivial computational task.

Non-likelihood marginal models are considered to be less sensitive to distributional assumptions, in particular about the association structure. In contrast, full likelihood methods require correct specification of the main effects as well as of the association structure to ensure consistency. Moreover, non-likelihood methods are usually easier to fit. Koch *et al.*⁴ suggested empirical generalized least squares. Unfortunately, this technique requires each covariate pattern to be non-sparse and hence continuous covariates lead to the breakdown of this method. Liang and Zeger⁵ proposed *generalized estimating equations* (GEEs). They differ from likelihood equations in that they only model the first moments (describing the marginal probabilities) of the joint distribution, and apply working assumptions to construct the information needed from the higher order moments. GEEs are regression models, and therefore easily handle continuous covariates. A drawback is that only when non-response is due to a *missing completely at random* mechanism⁶ do GEEs retain their consistency. For more general missing data mechanisms, special versions of GEEs have been developed.⁷

In many studies, questions about the marginal parameters are not the only ones of scientific interest, whence also the association parameters and/or the union probabilities are to be estimated. A few examples: 'What is the probability of failing on all outcomes?', 'What is the probability of having at least moderate relief during three visits in row?', 'What is the probability

for two raters to classify an individual in the same category or into “close” categories?”. In such cases a full likelihood analysis seems necessary.

Several full likelihood methods that directly model the marginal mean functions have been proposed. Bahadur⁸ proposed a marginal model which combines classical logistic regression for the marginal responses with two-way and higher order correlation coefficients to capture the association. Cox⁹ introduced an exponential family model for multivariate binary data, the parameters of which are interpreted as conditional logits, conditional log-odds ratios etc. While this model is easy to fit, it suffers from interpretational difficulties, since most often scientific questions are formulated in terms of marginal rather than conditional quantities. Compromises between full likelihood and GEE as introduced by Liang and Zeger⁵ have been proposed by Prentice,¹⁰ Zhao and Prentice,¹¹ Zhao *et al.*¹² and Liang *et al.*² These families, broadly referred to as GEE2, do not confine modelling to the marginal means but include the pairwise association structure (using correlations or odds ratios) in the modelling process, thereby still avoiding the need to model higher order association parameters. Fitzmaurice and Laird¹³ construct a model which combines the computational ease of a conditional model (Cox⁹) with the interpretational convenience of a marginal model. Because the parameter vector consists partly of marginal and partly of conditional parameters, it is called a mixed marginal-conditional parameterization in Fitzmaurice *et al.*¹⁴ Kauermann¹⁵ considers a similar approach, based on the work of Barndorff-Nielsen.¹⁶ Ekholm¹⁷ builds a marginal model purely using logit links. Precisely, logit links are assumed for the probability of a success, the probability of two successes simultaneously etc.

Ashford and Sowden³ suggested considering a vector of ordinal variables as a discrete realization of a multivariate normal variable. Their approach is known as the multivariate probit model. Dale¹⁸ defined the bivariate global odds ratio model, combining logit links for the marginal probabilities at each of two occasions with odds ratios to quantify the association. She resorts to the bivariate Plackett distribution¹⁹ to compute the joint probabilities required to fit her model. The Plackett distribution is also used in solving GEEs when the odds ratio is used to measure the association.²⁰ Several extensions of the Dale model from bivariate to multivariate outcomes have been proposed. Molenberghs and Lesaffre²¹ generalized the computations of the bivariate Plackett distribution in order to establish the multivariate cell probabilities. Their method involves solving polynomials of high degree and computing the derivatives thereof. McCullagh and Nelder²² defined a generalized linear model that incorporates the model of Molenberghs and Lesaffre in the case of a logit link for the marginal mean functions. They wrote the link function in terms of the joint probabilities, $X\beta = \eta = C \ln(A\mu)$, with X a design matrix, μ the vector of joint probabilities, A a matrix consisting solely of zeros and ones, so that $A\mu$ contains the marginal probabilities of all orders: the probabilities of each outcome separately, the probabilities for the cross-classification of all pairs of outcomes, for all triples, etc. Contrasts of log-probabilities are equated to a vector of linear predictors η using the contrast matrix C (of which elements are either 0, 1 or -1). Contrasts of log-probabilities encompass many commonly used links for both marginal probabilities and associations. Within this model formulation, the marginal means can be modelled via baseline-category logits, adjacent category logits, continuation ratio logits, or cumulative logits. The association can be described in terms of, for example, local or global odds ratios. As counterexamples, modelling the marginal distribution via the probit or the complementary log-log link is excluded.

Still, a marginal model is only useful if an efficient way to compute the joint probabilities is available. Besides the multivariate Plackett probabilities, Lang and Agresti²³ and Balagtas *et*

*al.*²⁴ consider the modelling process as equivalent to imposing restrictions on the multinomial probabilities. Hence, fitting can be done using undetermined Lagrange multiplier.²⁵ This method seems to work best in cases with a restricted set of covariate levels. Alternatively, the cell probabilities can be fitted using a Newton iteration scheme, as suggested by Glonek and McCullagh.²⁶ In this paper, we present a simple generalized linear model formulation for marginal modelling of multivariate categorical data. Different types of categorical data and a wide class of marginal and association models fit in the presented framework. For a detailed model presentation, we focus on ordinal data and prefer to describe the association in terms of marginal global odds ratios. An appealing feature of this formulation is that a series of seemingly *ad hoc* proposals for handling multivariate categorical data, such as the odds ratio model, the probit model and the Bahadur model, are unified.

We propose an easy to implement and fast fitting algorithm, avoiding the use of higher order polynomials (that lead to numerical problems for high dimensional contingency tables). It can be viewed as an adaptation of the iterative proportional fitting algorithm.²⁷ It is an efficient and stable tool to determine the joint probabilities when the association is in terms of marginal odds ratios. Its advantage over undetermined Lagrange multipliers²⁸ is that the dimensionality of the parameter vector does not increase with the number of covariate levels. In contrast to the high dimensional polynomials that need to be solved to determine the Plackett probabilities,²¹ the IPF seems to enjoy good numerical stability properties. Note that the IPF was also applied by Fitzmaurice and Laird¹³ with conditional odds ratios for binary data. Here, we show its usefulness with (global or local) marginal odds ratios.

Section 2 is devoted to the generic model formulation, while parameter estimation is discussed in Section 3. The psychiatric study is analysed in Section 4. A contingency table analysis is presented in Section 5.

2. MODEL FORMULATION

For each individual, subject, or experimental unit in a study, a series of measurements Y is recorded, together with covariate information x . The notation is as follows.

Let $i = 1, \dots, N$ indicate the covariate (design) level, containing n_i subjects. The outcome for subject r in the i th level (group) is a series of measurements $Y_{irt} (t = 1, \dots, T_i)$. Assume that variable Y_{irt} can take on c_i distinct (possibly ordered) values. Without loss of generality, denote them by $1, \dots, c_i$. All information about the responses on the units in the i th group is contained in a cross-classification of the outcomes Y_{irt} into a $c_1 \times \dots \times c_{T_i}$ dimensional contingency table with cell counts

$$Z_i^*(\mathbf{k}) \equiv Z_i^*(k_1, \dots, k_{T_i}). \quad (1)$$

Along with the outcomes, a vector of explanatory variables x_{it} is recorded. The covariate vector is allowed to change over time. It can include continuous and discrete variables. Available covariate information, along with other relevant design features, are incorporated in the design matrix X_i , further discussed in Appendix I.

In harmony with the possibility of using cumulative measures, construct the table of cumulative counts:

$$Z_i(\mathbf{k}) = \sum_{l \leq \mathbf{k}} Z_i^*(l). \quad (2)$$

Thus, $Z_i(\mathbf{k})$, where $\mathbf{k} = (k_1, \dots, k_{T_i})$, is just the number of individuals in group i whose observed response vector is \mathbf{k} , and likewise for $Z_i(\mathbf{k})^*$. The corresponding probabilities are

$$\mu_i^*(\mathbf{k}) = \text{pr}(Y_{ir} = \mathbf{k} | X_i, \boldsymbol{\beta}) \quad (3)$$

and $\mu_i(\mathbf{k}) = \text{pr}(Y_{ir} \leq \mathbf{k} | X_i, \boldsymbol{\beta})$. Let \mathbf{Z}_i be the vector of all cumulative cell counts with $\boldsymbol{\mu}_i$ the corresponding vector of probabilities. Note that $Z_i(c_1, \dots, c_{T_i}) = n_i$ and $\mu_i(c_1, \dots, c_{T_i}) = 1$. Therefore, omitting these two entries from \mathbf{Z}_i and $\boldsymbol{\mu}_i$, respectively, yields non-redundant sets. Similarly, \mathbf{Z}_i^* and $\boldsymbol{\mu}_i^*$ are defined, and simple matrix equalities

$$\boldsymbol{\mu}_i^* = B_i \boldsymbol{\mu}_i, \quad \mathbf{Z}_i^* = B_i \mathbf{Z}_i \quad (4)$$

hold. As an example, consider a bivariate binary outcome vector, with counts $\boldsymbol{\mu}_i^* = (\mu_{11}^*, \mu_{12}^*, \mu_{21}^*, \mu_{22}^*)$ and a similar ordering of $\boldsymbol{\mu}_i$. The matrix B_i is found by

$$B_i^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The marginal counts are given by all counts for which all but one indexes are equal to their maximal value: $Z_{irk} \equiv Z_i(c_1, \dots, c_{t-1}, k, c_{t+1}, \dots, c_{T_i})$. Bivariate cell counts, that is, cell counts of a cross-classification of a pair of outcomes, follow from setting all but two indexes k_s equal to c_s . Therefore, this description very naturally combines univariate, bivariate and multivariate information. The ordering needed to stack the multi-indexed counts and probabilities into a vector will be assumed fixed. Several orderings of both \mathbf{Z}_i and $\boldsymbol{\mu}_i$ are possible. A natural choice is the lexicographic ordering, but this has the disadvantage of dispersing the univariate marginal counts and means over the entire vector. Therefore, we will group the elements first by dimensionality.

2.1. Choices of Link Functions

For the vector of links $\boldsymbol{\eta}_i$ we consider a function, mapping the C_i -vector $\boldsymbol{\mu}_i$ ($C_i = c_1, c_2, \dots, c_{T_i}$) to

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_i(\boldsymbol{\mu}_i), \quad (5)$$

a C_i -vector. Often, $C_i = C_i'$, and $\boldsymbol{\eta}_i$ and $\boldsymbol{\mu}_i$ have the same ordering. A counterexample is provided by the probit model, where the number of link functions is smaller than the number of mean components, as soon as $T_i > 2$ (see (13)–(15)). As already indicated in the introduction, an important class of link functions is discussed by McCullagh and Nelder:²²

$$\boldsymbol{\eta}_i(\boldsymbol{\mu}_i) = C \ln(A\boldsymbol{\mu}_i) \quad (6)$$

a definition in terms of contrasts of log probabilities, where the probabilities involved are linear combination $A\boldsymbol{\mu}$.

We consider particular choices of link functions. Let us abbreviate the univariate marginal probabilities by $\mu_{itk} = \mu_i(c_1, \dots, c_{t-1}, \mathbf{k}, c_{t+1}, \dots, c_{T_i})$, then the logit link becomes $\eta_{itk} = \ln(\mu_{itk}) - \ln(1 - \mu_{itk}) = \text{logit}(\mu_{itk})$. Some link functions that are occasionally of interest, such as the probit or complementary log-log link, are not supported by (6). They can easily be included in (5). The probit link is $\eta_{itk} = \Phi_1^{-1}(\mu_{itk})$, with Φ_1 the univariate standard normal distribution.

However, univariate links alone do not fully specify $\boldsymbol{\eta}_i$, and hence leave the joint distribution partly undetermined. Full specification of the association requires addressing the form of pairwise and higher-order probabilities. First, we will consider the pairwise associations. Let us denote the bivariate probabilities, pertaining to the t th and s th outcomes, by

$$\mu_{i,ts,kl} = \mu_i(c_1, \dots, c_{t-1}, \mathbf{k}, c_{t+1}, \dots, c_{s-1}, l, c_{s+1}, \dots, c_{T_i}).$$

Some association parameterizations are summarized in Table I.

The success probability parameterization of Ekholm¹⁷ consists of choosing a link function for the univariate marginal means (for example, a logit link) and then applying the same link function to the two and higher order success probabilities (that is, the probabilities for observing a single success when looking at one outcome at a time, a pair of successes when looking at pairs of outcomes, ...). For categorical data, a logit link for two-way probabilities is given by

$$\eta_{i,ts,kl} = \ln(\mu_{i,ts,kl}) - \ln(1 - \mu_{i,ts,kl}) = \text{logit}(\mu_{i,ts,kl}) \quad (7)$$

for $k = 1, \dots, c_t - 1$ and $l = 1, \dots, c_s - 1$. The marginal correlation coefficient⁸ is defined as

$$\rho_{i,ts,kl} = \frac{\mu_{i,ts,kl} - \mu_{itk}\mu_{isl}}{\sqrt{\{\mu_{itk}(1 - \mu_{itk})\mu_{isl}(1 - \mu_{isl})\}}}. \quad (8)$$

It is convenient to equate the corresponding element in the link function to Fisher's z -transform of $\rho_{i,ts,kl}$ or a simple function of it: $\eta_{i,ts,kl} = \ln(1 + \rho_{i,ts,kl}) - \ln(1 - \rho_{i,ts,kl})$. Higher order 'correlations' are defined in terms of standardized cumulants. This definition extended the use of the marginal correlations from multivariate binary data (its classical use) to multivariate ordinal data.

We will mainly be concerned with the marginal global odds ratio, defined by

$$\psi_{i,ts,kl} = \frac{(\mu_{i,ts,kl})(1 - \mu_{itk} - \mu_{isl} + \mu_{i,ts,kl})}{(\mu_{isl} - \mu_{i,ts,kl})(\mu_{itk} - \mu_{i,ts,kl})} \quad (9)$$

and usefully modelled on the log scale as

$$\eta_{i,ts,kl} = \ln(\mu_{i,ts,kl}) - \ln(\mu_{itk} - \mu_{i,ts,kl}) - \ln(\mu_{isl} - \mu_{i,ts,kl}) + \ln(1 - \mu_{itk} - \mu_{isl} + \mu_{i,ts,kl}).$$

Higher order global odds ratios are easily introduced using ratios of conditional odds (ratios). Let

$$\mu_{it|s}(z_s) = \text{pr}(Z_{irt_t} = 1 | Z_{irs_s} = z_s, X_i, \beta) \quad (10)$$

be the conditional probability of observing a success at occasion t , given the value z_s is observed at occasion s , and write the corresponding conditional odds as $\psi_{it|s}(z_s) = \mu_{it|s}(z_s)/1 - \mu_{it|s}(z_s)$. The

Table I. Association structure of selected marginal models

Name	Association structure	Equation
Success probability	Logit of joint probability	(7)
Bahadur model	Marginal correlation coefficients	(8)
Dale model	Global marginal odds ratio	(9)–(11)
	Local marginal odds ratio	(12)
Probit model	Polychoric correlation	(13)–(15)

pairwise marginal odds ratio, for occasions t and s , is defined as

$$\psi_{its} = \frac{\{\text{pr}(Z_{irtk_t} = 1, Z_{irsk_s} = 1)\} \{\text{pr}(Z_{irtk_t} = 0, Z_{irsk_s} = 0)\}}{\{\text{pr}(Z_{irtk_t} = 0, Z_{irsk_s} = 1)\} \{\text{pr}(Z_{irtk_t} = 1, Z_{irsk_s} = 0)\}} = \frac{\psi_{it|s}(1)}{\psi_{it|s}(0)}$$

in accordance with (9). This formulation can be exploited to define the higher order marginal odds ratios in a recursive fashion:

$$\psi_{it_1 \dots t_m t_{m+1}} = \frac{\psi_{it_1 \dots t_m t_{m+1}}(1)}{\psi_{it_1 \dots t_m t_{m+1}}(0)} \tag{11}$$

where $\psi_{it_1 \dots t_m t_{m+1}}(z_{m+1})$ is defined by conditioning all probabilities occurring in the expression for $\psi_{it_1 \dots t_m}$ on $Z_{irt_{m+1}} = z_{m+1}$. The choice of the variable to condition on is immaterial. Observe that multi-way marginal global odds ratios are defined solely in terms of conditional probabilities.

Another type of marginal odds ratios are the marginal *local* odds ratios. This changes (9) to

$$\psi_{i,ts,kl}^* = \frac{\mu_{i,ts,kl}^* \mu_{i,ts,k+1,l+1}^*}{\mu_{i,ts,k+1,l}^* \mu_{i,ts,k,l+1}^*} \tag{12}$$

with the cell probabilities as in (3). Higher order marginal local odds ratios are constructed in the same way as their global counterparts. Various types of odds ratios will be discussed in Section 3.2.

Observe that the multivariate probit model³ also fits within the class defined by (5). Let $g = h^{-1}$. For three categorical outcome variables, the inverse link is specified by

$$\mu_{itk} = \Phi_1(\eta_{itk}) \tag{13}$$

$$\mu_{i,ts,kl} = \Phi_2(\eta_{itk}, \eta_{isb}, \eta_{i,ts,kl}) \tag{14}$$

$$\mu_{i,123,klm} = \Phi_3(\eta_{i1k}, \eta_{i2l}, \eta_{i3m}, \eta_{i,12,kl}, \eta_{i,13,km}, \eta_{i,23,lm}) \tag{15}$$

where the notation for the three-way probabilities is obvious. The association links $\eta_{i,ts,kl}$ represent any transform (for example, Fisher’s z-transform) of the polychoric correlation coefficient. It is common practice to keep each correlation constant throughout a table, rather than having it depend on the categories: $\eta_{i,ts,kl} \equiv \eta_{i,ts}$. Relaxing this requirement may still give a valid set of probabilities, but the correspondence between the categorical variables and a latent multivariate normal variable is lost. Finally, observe that univariate links and bivariate links (representing correlations) fully determine the joint distribution. This implies that the mean vector and the link vector will have different length, except in the univariate and bivariate cases.

In summary, marginal models are characterized by jointly specifying marginal response functions and marginal association measures. Models can be classified by the association measures, as exemplified in Table I.

Finally, model formulation is completed by constructing appropriate design matrices. An example is given in Appendix I.

3. MAXIMUM LIKELIHOOD ESTIMATION

We first discuss the form of the likelihood equations and introduce algorithms to obtain the maximum likelihood estimator, as well as a way to estimate its precision. Then, the algorithm to determine the joint probabilities is presented.

3.1. Score equations and maximization

Under a multinomial sampling scheme, the kernel of the log-likelihood, in terms of the observations \mathbf{Z}_i^* and the corresponding cell probabilities $\boldsymbol{\mu}_i^*$, is

$$l(\boldsymbol{\beta}; \mathbf{Z}^*) = \sum_{i=1}^N \mathbf{Z}_i^{*\text{T}} \ln(\boldsymbol{\mu}_i^*(\boldsymbol{\beta})).$$

When working with the cumulative counts \mathbf{Z}_i and the cumulative probabilities $\boldsymbol{\mu}_i$, and knowing that relations (4) hold, we can rewrite the log-likelihood as

$$l(\boldsymbol{\beta}; \mathbf{Z}) = \sum_{i=1}^N (\mathbf{B}_i \mathbf{Z}_i)^{\text{T}} \ln(\mathbf{B}_i \boldsymbol{\mu}_i(\boldsymbol{\beta})). \quad (16)$$

The derivative of the contribution of group i to (16) with respect to $\boldsymbol{\mu}_i$ is then given by

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\mu}_i} &= \{\mathbf{B}_i^{\text{T}} [\text{diag}(\boldsymbol{\mu}_i^*)]^{-1} \mathbf{B}_i\} (\mathbf{Z}_i - n_i \boldsymbol{\mu}_i) \\ &= \{\mathbf{B}_i^{\text{T}} \text{cov}(\mathbf{Z}_i^*)^{-1} \mathbf{B}_i\} (\mathbf{Z}_i - n_i \boldsymbol{\mu}_i) \\ &= \text{cov}(\mathbf{Z}_i)^{-1} (\mathbf{Z}_i - n_i \boldsymbol{\mu}_i). \end{aligned} \quad (17)$$

Given (17), the score function becomes

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \right)^{\text{T}} \left[\left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\mu}_i} \right)^{\text{T}} \right]^{-1} V_i^{-1} \mathbf{S}_i \quad (18)$$

with $\mathbf{S}_i = \mathbf{Z}_i - n_i \boldsymbol{\mu}_i$, and $V_i = \text{cov}(\mathbf{Z}_i)$. A typical element of V_i is

$$\text{cov}(z_i(k_1 \dots k_{T_i}), z_i(l_1, \dots, l_{T_i})) = \mu_i(m_1, \dots, m_{T_i}) - \mu_i(k_1, \dots, k_{T_i}) \cdot \mu_i(l_1, \dots, l_{T_i})$$

where $m_t = \min(k_t, l_t)$.

Computation of the matrix $\mathbf{Q}_i = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\mu}_i$ is extremely simple if the link is of the form (6), because then²⁹

$$\mathbf{Q}_i = \mathbf{C} \{ \text{diag}(\mathbf{A} \boldsymbol{\mu}_i) \}^{-1} \mathbf{A}. \quad (19)$$

This motivates our choice to compute \mathbf{Q}_i and invert it, rather than computing \mathbf{Q}_i^{-1} directly, as was done by Molenberghs and Lesaffre.²¹

Replacing the univariate marginal link functions in (6), $\boldsymbol{\eta}_i^{(1)}$ say, by any other inverse cumulative distribution function F^{-1} with probability density function f , and retaining the specification of

the association in terms of a form satisfying (6), yields the expression

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\mu}) = \begin{pmatrix} F^{-1}(\boldsymbol{\mu}^{(1)}) \\ C_2 \ln(A\boldsymbol{\mu}) \end{pmatrix}$$

with corresponding derivative

$$Q_i = \begin{pmatrix} \text{diag}\{f(\boldsymbol{\eta}^{(1)})\}^{-1}; & 0 \\ C_2 \{\text{diag}(A\boldsymbol{\mu})\}^{-1}A & \end{pmatrix}. \tag{20}$$

The matrix C_2 is similar to the matrix C in (6), but now only applies to the association part of the model. Choosing $F = \Phi$ and $f = \phi$ (the standard normal distribution and density functions), we obtain a global odds ratio model with univariate probit links.

As discussed in the previous section, the multivariate probit model also fits within the proposed framework. In this case, one prefers to compute the matrix Q_i^{-1} , rather than its inverse, unlike with the global odds ratio model, or any other model of the form (6). Although in the probit case the matrix Q_i^{-1} is easier to compute than Q_i , the computations are still more complex than calculating (20). The components are the derivatives of multivariate standard normal distribution functions. The evaluation of multivariate normal integrals is required. Lesaffre and Molenberghs³⁰ chose to use the algorithm proposed by Shervish.³¹ In the common case of linear predictors, the derivative of the link vector with respect to $\boldsymbol{\beta}$ is the design matrix X_i .

The maximum likelihood estimator satisfies $\mathbf{U}(\hat{\boldsymbol{\beta}}) = 0$. Two popular fitting algorithms are Fisher scoring and the Newton–Raphson algorithm. In the case of Fisher scoring, one starts with a vector of initial estimates $\boldsymbol{\beta}^{(0)}$ and updates the current value of the parameter vector $\boldsymbol{\beta}^{(t)}$ by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + W(\boldsymbol{\beta}^{(t)})^{-1}U(\boldsymbol{\beta}^{(t)}) \tag{21}$$

with

$$W(\boldsymbol{\beta}) = \sum_{i=1}^N n_i \begin{pmatrix} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \end{pmatrix}^T \left[\begin{pmatrix} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\mu}_i} \end{pmatrix}^T \right]^{-1} V_i^{-1} \left[\begin{pmatrix} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\mu}} \end{pmatrix} \right]^{-1} \begin{pmatrix} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \end{pmatrix}.$$

The expected information matrix assumes the form $W(\boldsymbol{\beta})$, estimated by $W(\hat{\boldsymbol{\beta}})$. A Newton–Raphson iteration scheme is found by substituting the matrix $W(\boldsymbol{\beta})$ in (21) by $H(\boldsymbol{\beta})$, the matrix of second derivatives of the log-likelihood. An outline of this procedure for cumulative counts is presented in Appendix II.

3.2. Determining the Joint Probabilities

In order to compute the score equations and to implement the updating algorithm, knowledge of the multivariate cumulative probabilities $\boldsymbol{\mu}_i$ is required. The choice of a fitting technique will strongly depend on the choice of link functions. For multivariate odds ratio models (also referred to as multivariate Dale models) several proposals have been made, such as the use of multivariate Plackett probabilities,²¹ the use of Lagrange multipliers,²³ and a Newton iteration mechanism.²⁶ With the Plackett probability approach, we found that for four and higher dimensional problems, the derivatives of high dimensional polynomials can become numerically unstable. Here, the iterative proportional fitting (IPF) algorithm is adapted to produce a quick and reliable tool to compute the cumulative probabilities. A similar use of the IPF algorithm was proposed by Kauermann.¹⁵ Owing to the use of score function (18), there is no need to compute the derivatives

of the probabilities directly since Q_i easily follows from (19), leaving only the probabilities to be computed.

Given the marginal probabilities and the odds ratio parameters, our IPF algorithm produces a multidimensional table of cell probabilities. The IPF algorithm is known from its use in fitting log-linear models,³² where the association is described using *conditional* odds ratios. The algorithm was also applied by Fitzmaurice and Laird¹³ for their mixed marginal-conditional models. In our fully marginal models, marginal odds ratios are used. We distinguish between two types. Global odds ratios, given in (9)–(11), are relevant for ordinal responses,³³ while local odds ratios, as in (12), are a natural choice for nominal outcomes. Of course, both sets coincide for binary responses.

We will describe our algorithm for global odds ratios first, and then discuss the local odds ratio version in the concluding paragraph of this section. We need to determine the cumulative probabilities $\mu_i(k_1, \dots, k_{T_i})$ which correspond to cumulative cell $Z_i(k_1, \dots, k_{T_i})$. Recall that this notation encompasses not only T_i -way classifications, but also one-way, two-way etc. classifications, by setting an appropriate set of indices $k_t = c_t$. Omitting indices for which $k_t = c_t$, we assume without loss of generality that we need to determine a K -way probability $\mu_i(k_1, \dots, k_K)$, with $k_t < c_t$ for all t .

We will proceed recursively. First, note that the cumulative probabilities $\mu_i(l_1, \dots, l_K)$, with $l_t \in \{k_t, c_t\}$ for $t = 1, \dots, K$, completely describe a 2^K contingency table. Secondly, as soon as at least one $l_t = c_t$, we obtain a lower order probability. Our recursion will be based on the assumption that these lower order probabilities have been determined. The starting point of the inductive construction is obtained by setting all but one $l_t = c_t$, whence we obtain univariate probabilities μ_{itk} , which are of course easy to determine from the marginal links η_{itk} . Drop the index i from notation.

From the cumulative probabilities, we determine the cell probabilities $\mu_{k_1 \dots k_K}^{j_1 \dots j_K}$, with $j_t = 1, 2$ and adopt the convention that the K -way cumulative cell probabilities are incorporated as

$$\mu_{k_1 \dots k_K}^{1 \dots 1} = \mu(k_1, \dots, k_K). \quad (22)$$

We will explicitly need the cell probabilities of dimension $K - 1$:

$$\mu_{k_1 \dots k_{t-1} k_{t+1} \dots k_K}^{j_1 \dots j_{t-1} j_{t+1} \dots j_K} = \sum_{j_t=1}^2 \mu_{k_1 \dots k_K}^{j_1 \dots j_K}.$$

The IPF algorithm is started by choosing a table of initial values:

$$\mu_{k_1 \dots k_K}^{j_1 \dots j_K}(0) = \begin{cases} \psi_i(k_1, \dots, k_K) & \text{if } (j_1, \dots, j_K) = (1, \dots, 1), \\ 1 & \text{else} \end{cases}$$

with $\psi_i(k_1, \dots, k_K) = \exp(\eta_i(k_1, \dots, k_K))$, the corresponding global odds ratio. The table clearly has the correct association structure, but the marginals are incorrect and the sum of the cell counts is not equal to one. Updating cycle $(m + 1)$ requires K substeps, to match each of the $K - 1$ dimensional marginal tables:

$$\mu_{k_1 \dots k_K}^{j_1 \dots j_K} \left(m + \frac{t}{K} \right) = \mu_{k_1 \dots k_K}^{j_1 \dots j_K} \left(m + \frac{t-1}{K} \right) \frac{\mu_{k_1 \dots k_{t-1} k_{t+1} \dots k_K}^{j_1 \dots j_{t-1} j_{t+1} \dots j_K}}{\mu_{k_1 \dots k_{t-1} k_{t+1} \dots k_K}^{j_1 \dots j_{t-1} j_{t+1} \dots j_K} \left(m + \frac{t-1}{K} \right)}$$

($t = 1, \dots, K$), the argument of μ indicating the iteration subcycle. Upon convergence, (22) can be used to identify the required K -way probability.

Convergence of the IPF algorithm is established in Csiszar.³⁴ However, the parameter space of the marginal odds ratios is constrained, unless in the special case of a constant odds ratio for a bivariate outcome.² A violation of the constraints will be revealed by a cumulative probability vector with negative entries. If this occurs in the course of an updating algorithm, appropriate action (for example, step halving) has to be taken. Note that the authors never encountered problems of this kind, suggesting that the constraints are very mild.

For marginal local odds ratios a slightly adapted and simpler procedure is proposed. Instead of considering subsets of binary variables, we now consider the whole marginal multi-way table directly. With a similar recursive argument, we assume that the full set of marginal tables up to dimension $K - 1$ is determined. Then, we construct a K -dimensional initial table

$$\mu_i^*(k_1, \dots, k_K)(0) = \prod_{c_i > l_i \geq k_i} \psi_i^*(l_1, \dots, l_K)$$

for all cells (k_1, \dots, k_K) . This table clearly has got the required K -way local association structure. The updating algorithm matches the entries to the K sets of $K - 1$ -dimensional marginal tables.

4. ANALYSIS OF PSYCHIATRIC STUDY

We analyse the cross-over and longitudinal parts of the influenza study in turn.

4.1. Cross-over study

We analyse presence/absence of headache (H) and presence/absence of respiratory problems (R), measured in both trial periods. We combine marginal logits with marginal log-odds ratios. The modelling is in stages. First, period effect is included. Then, a contrast between the two companies, a contrast between the two vaccinations, and an interaction term between companies and treatments is added. Further, the baseline covariates AGE (in years) and SEX (0 = male, 1 = female) are included. There are three types of two-way association: between the two headache outcomes; between the two respiratory problems outcomes; and between a headache and a respiratory outcome. The two-way associations are graphically depicted in Figure 1. Three-way and four-way associations are assumed to be constant throughout. The results are presented in Table II.

A positive parameter decreases the odds for headache/respiratory problems. Respiratory problems are on average very infrequent, as can be seen from the high value of the intercept. For both outcomes, there is a significant period effect; there are fewer headaches and respiratory problems in the second period. Also, the influenza vaccination causes fewer headaches, but more respiratory problems. Headaches are more frequently seen in younger people, while the opposite holds for respiratory problems. Men suffer more from headaches after vaccination than women. The odds ratio between two respiratory problems is high (7.9), while a somewhat smaller association is seen between the pair of headache measures (3.1) and between the mixed pair (3.0). This is due to the fact that respiratory problems are more severe and probably more strongly related with vaccination than headache, which can have various causes. Extending the two-way association structure to include a company A effect was not significant. We found no higher order association.

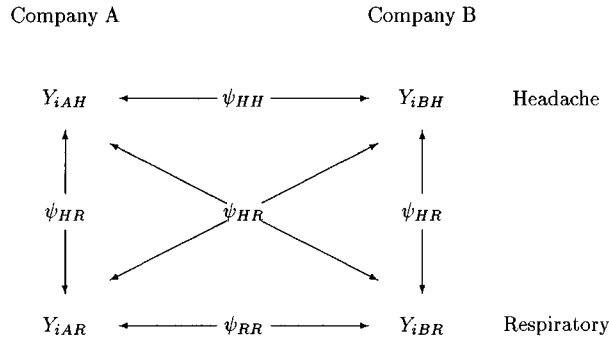


Figure 1. Association structure for influenza cross-over study

Table II. Parameter estimates (standard errors) for the influenza data cross-over trial

Parameters for headache	
Intercept	0.055 (1.092)
Period effect	0.434 (0.140)
Company A effect	-0.341 (0.221)
Influenza effect	0.132 (0.212)
Company A-influenza interaction	-0.053 (0.281)
Age	0.052 (0.054)
Sex	0.875 (0.217)
<i>Respiratory problems</i>	
Intercept	5.217 (1.297)
Period effect	0.167 (0.156)
Company A effect	-0.229 (0.267)
Influenza effect	-0.119 (0.226)
Company-influenza interaction	0.257 (0.312)
Age	-0.159 (0.063)
Sex	0.133 (0.243)
<i>Association (log-odds ratios)</i>	
Headache-headache	1.130 (0.251)
Respiratory-respiratory	2.061 (0.309)
Headache-respiratory	1.090 (0.191)
Three-way interaction	0.219 (0.395)
Four-way interaction	2.822 (1.462)

4.2. The longitudinal study

Pain was measured on six consecutive days after vaccination. Changes in response are mainly observed during the first three days. Significant predictors for the evolution of pain level are SEX, AGE, the use of medication (MED), and the actual vaccination. The effect of all covariates is

Table III. Parameter Estimates (Standard Errors) for the Influenza Longitudinal Data

Parameter	Marginal parameters		
	average	linear	quadratic
Intercept 1	- 2.344 (0.995)	1.236 (0.928)	- 0.599 (0.361)
Intercept 2	0.340 (0.995)	0.893 (0.926)	- 0.803 (0.365)
Age	0.147 (0.049)	- 0.013 (0.046)	0.047 (0.018)
Sex	0.427 (0.185)	- 0.307 (0.169)	- 0.014 (0.066)
Medication	- 0.469 (0.216)	- 0.256 (0.186)	0.074 (0.079)
Company A effect	1.226 (0.274)	0.367 (0.270)	- 0.393 (0.105)
Influenza effect	- 0.739 (0.213)	0.082 (0.192)	- 0.111 (0.074)
Company-influenza interaction	- 1.055 (0.336)	- 0.259 (0.316)	0.343 (0.124)
<i>Associations (log-odds ratios)</i>			
Time 1-time 2	1.809 (0.208)		
Time 1-time 3	0.976 (0.258)		
Time 2-time 3	3.402 (0.405)		
3-way interaction	0.878 (0.625)		

allowed to change over time. As there are four vaccinations, we decompose them into two factors (company A, influenza, and the interaction). At each measurement time, there are two intercepts, corresponding to two cumulative logits (no pain (0) versus pain (1 and 2); no or mild pain (0 and 1) versus moderate pain (2)). All covariates are allowed to have a different effect at each measurement, presented as SEX(overall), SEX(linear), and SEX(quadratic). The results are presented in Table III. We observe strong quadratic time effects for company A and for the interaction between company A and influenza. Considering the hepatitis vaccine for company B as the baseline, the differences (for each measurement time) on the logit scale between each vaccine and the baseline are: for the influenza vaccine of company A, - 5.33, 0.85 and - 1.10; for the influenza vaccine of company B, - 1.36, 1.95 and 0.43; for the hepatitis vaccine of company A, - 4.18, - 1.15 and - 1.71. The combination of a strong interaction between company and type of vaccine and of the change of the effects over time, yields a complex picture. As the outcomes are modelled via marginal logits, they are interpreted using standard logistic regression methodology. Making comparisons for the three measurement times, we are able to study the evolution of differences over time.

5. A LOCAL CROSS-RATIO MODEL

Goodman³⁵ studied association models for two-way contingency tables with ordered categories. The cross-classification of eye colour and hair colour of 5387 children is reproduced in Table IV. Goodman treated these responses as ordinal which, although sensible, might be open to discussion. His association models were conditional in nature but arguably marginal models are easier to interpret. We combine marginal probabilities, one set for each variable, with local odds ratios to describe the association. We consider two models. The first one (8 parameters) assumes a constant local odds ratio. The simpler model which assumes independence between both responses has been shown by Goodman to provide a poor fit and will not be considered here. The second, saturated, model allows an unstructured 3×4 table of local odds ratios. The marginal

Table IV. Eye colour and hair colour of 5387 children in Caithness (Goodman³⁵)

Eye colour	Hair colour				
	Fair	Red	Medium	Dark	Black
Blue	326	38	241	110	3
Light	688	116	584	188	4
Medium	343	84	909	412	26
Dark	98	48	403	681	85

probabilities for *both* models are (0.13, 0.29, 0.33, 0.25) for eye colour and (0.27, 0.05, 0.40, 0.26, 0.02) for hair colour. The common local odds ratio for the first model equals 1.50. The deviance is 131.10 on 11 degrees of freedom, rejecting the constant (or uniform) association model. Note that Goodman's *conditional* model for uniform association exhibited a much poorer deviance of 265.03 on 11 degrees of freedom. The 12 local odds ratios, organized as an association table, are:

	1/2	2/3	3/4	4/5
1/2	1.45	0.79	0.71	0.78
2/3	1.45	2.15	1.41	2.97
3/4	2.00	0.78	3.73	1.98

6. CONCLUDING REMARKS

We have presented a general framework to construct marginal models for multivariate categorical data. Although we focused on ordinal data, we also indicated how nominal outcomes can be analysed. A wide range of models for the responses, as well as measures for the association between the responses, can be incorporated in the model. We emphasized the use of marginal global odds ratios. The approach allows for the incorporation of covariates and a wide class of designs (clustering, longitudinal data, cross-over trials, etc.).

For the special case where the association is described in terms of marginal odds ratios, we developed a flexible, easy and stable fitting algorithm, using the iterative proportional fitting technique. The advantage over other implementations is that we are able to fit models with a relatively large number of outcomes. We encountered no problems for about seven repeated measurements. Still, the exponentially increasing number of parameters in full likelihood models will always impose constraints on the feasibility. Arguably, for very long series of measurements, non-likelihood modes of inference need to be sought. Although not our primary goal, our approach also results in computation time gain, as we only need to determine the probabilities, avoiding the direct computation of their derivatives. The relative time gain increases with the number of categories and cutpoints. For a model with three ordinal outcomes on a 5-point scale, our approach is 15 times faster than the Plackett probabilities method. Further, fitting models with continuous covariates poses no problems.

Model specification can be investigated by goodness-of-fit methodology, especially in the case of a bounded number of possible covariate levels. Person's X^2 or deviance G^2 statistics can be

used. Also, empirical corrections to routine inferential procedures as advocated by White³⁶ are also possible.

Problems involving missing data can be tackled with the same ease as described in Molenberghs and Lesaffre²¹ and Kenward *et al.*³⁷ Indeed, an attractive property of fully marginal models is their reproducibility.² Even if the number of outcomes is *by design* the same for each unit, rendering a conditional approach plausible in theory, this balance is often destroyed due to missing data. Fully marginal models can easily deal with this type of unbalanced data, whether arising by design or unplanned missingness. An example of such an approach is given in Molenberghs *et al.*³⁸

APPENDIX I: DESIGN

A complete model description requires the introduction of a design matrix. It will be indicated how model assumptions are reflected by choosing particular types of design. We deliberately restrict ourselves to linear predictors, while, in principle, there is no obstacle to include non-linear effects. The design matrix X_i for the i th individual includes all information which is needed to model both the marginal mean functions and associations. Each row corresponds to an element in the vector of link functions η_i . Its generality is best illustrated using an example.

Consider the case of three outcomes, recorded on a three-point scale. Let the measurement times be $t_1 \equiv 0$, t_2 and t_3 . Assume the recording of four explanatory variables, x_1, \dots, x_4 , with only x_3 and x_4 time-varying. We first turn our attention to the marginal distributions. Let x_1 have a constant effect on each outcome, that is, a single parameter describes the effect of x_1 on the cumulative logits of the three outcome probabilities. On the other hand, the effect of x_2 is allowed to change over time. We also introduce a single parameter to describe the effect of x_3 and three separate parameters to account for the influence of x_4 . These assumptions call for the following parameter vector:

$$\beta_1 = (\alpha_1, \alpha_2, \tau_2, \tau_3, \beta_1, \beta_{21}, \beta_{22}, \beta_{23}, \beta_3, \beta_{41}, \beta_{42}, \beta_{43})^T$$

with α_k intercepts, τ_t the effect of measurement time t , β_1 and β_3 the parameters needed to describe the effect of x_1 and x_3 , respectively, and β_{jt} the parameter describing the effect of $x_j^{(t)}$ at time t ($j = 2, 4$; $t = 1, 2, 3$). Next, assume that the two-way associations depend on the pair of variables they refer to, as well as on the cumulative category within that variable. Finally, assume dependence on the covariate x_{1i} . This introduces extra parameters

$$\beta_2 = (\gamma, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \gamma_{31}, \gamma_{32}, \phi_1, \phi_2, \phi_3)^T$$

with γ the intercept, γ_{tk} the dependence on category k of outcome t ($t = 1, 2$; $k = 1, 2$), and ϕ_t the dependence on x_1 . Finally, assume a constant value for the three-way associations, β_3 say. The entire parameter vector is denoted as $\beta = (\beta_1^T, \beta_2^T, \beta_3)^T$. The design matrix for subject i , $X_{i\cdot}$ is block diagonal with blocks X_{i1} (mean functions, shown in Figure 2(a)), X_{i2} (pairwise association, shown Figure 2(b)) and X_{i3} (three-way association).

Observe that, apart from the intercepts α_j , the design is identical for each cumulative logit in Figure 2(a). This reflects the proportional odds assumption when marginal logits are used. If this assumption is considered unrealistic, the design can be generalized without any difficulty. Nominal covariates and interactions between covariates are also easily included.

The second block of the design matrix, X_{2i} , corresponds to the pairwise associations and is given by Figure 2(b). Finally, the design for the three-way associations in our example is a 8×1

a) Marginal Means

	α_1	α_2	τ_2	τ_3	β_1	β_{21}	β_{22}	β_{23}	β_3	β_{41}	β_{42}	β_{43}
$\eta_i(1, 3, 3)$	1	0	0	0	x_{1i}	x_{2i}	0	0	$x_{3i}^{(1)}$	$x_{4i}^{(1)}$	0	0
$\eta_i(2, 3, 3)$	0	1	0	0	x_{1i}	x_{2i}	0	0	$x_{3i}^{(1)}$	$x_{4i}^{(1)}$	0	0
$\eta_i(3, 1, 3)$	1	0	1	0	x_{1i}	0	x_{2i}	0	$x_{3i}^{(2)}$	0	$x_{4i}^{(2)}$	0
$\eta_i(3, 2, 3)$	0	1	1	0	x_{1i}	0	x_{2i}	0	$x_{3i}^{(2)}$	0	$x_{4i}^{(2)}$	0
$\eta_i(3, 3, 1)$	1	0	0	1	x_{1i}	0	0	x_{2i}	$x_{3i}^{(3)}$	0	0	$x_{4i}^{(3)}$
$\eta_i(3, 3, 2)$	0	1	0	1	x_{1i}	0	0	x_{2i}	$x_{3i}^{(3)}$	0	0	$x_{4i}^{(3)}$

b) Pairwise Associations

	γ	γ_{11}	γ_{12}	γ_{21}	γ_{22}	γ_{31}	γ_{32}	ϕ_1	ϕ_2	ϕ_3
$\eta_i(1, 1, 3)$	1	1	0	1	0	0	0	x_{1i}	0	0
$\eta_i(1, 2, 3)$	1	1	0	0	1	0	0	x_{1i}	0	0
$\eta_i(2, 1, 3)$	1	0	1	1	0	0	0	x_{1i}	0	0
$\eta_i(2, 2, 3)$	1	0	1	0	1	0	0	x_{1i}	0	0
$\eta_i(1, 3, 1)$	1	1	0	0	0	1	0	0	x_{1i}	0
$\eta_i(1, 3, 2)$	1	1	0	0	0	0	1	0	x_{1i}	0
$\eta_i(2, 3, 1)$	1	0	1	0	0	1	0	0	x_{1i}	0
$\eta_i(2, 3, 2)$	1	0	1	0	0	0	1	0	x_{1i}	0
$\eta_i(3, 1, 1)$	1	0	0	1	0	1	0	0	0	x_{1i}
$\eta_i(3, 1, 2)$	1	0	0	1	0	0	1	0	0	x_{1i}
$\eta_i(3, 2, 1)$	1	0	0	0	1	1	0	0	0	x_{1i}
$\eta_i(3, 2, 2)$	1	0	0	0	1	0	1	0	0	x_{1i}

Figure 2. Design matrix for marginal means and pairwise associations

column vector of ones, corresponding to the eight link functions $\eta_i(k_1, k_2, k_3)$ ($k_t = 1, 2; t = 1, 2, 3$). Replacing the elements of this vector by zeros has the effect of setting higher order association components equal to one (zero on the log scale).

Generalizations include non-block diagonal designs, and structured association such as exchangeable association, temporal association (as introduced by Fitzmaurice and Lipsitz³⁹), and banded association. In many circumstances, the association structure of a given table, representing a two- or multi-way classification of several variables is of direct interest, rather than the dependence of the outcomes on covariates. Association measures are extensively studied in Goodman.³⁵ With our approach, we are also able to explore the association structure of contingency tables. A typical form for the linear predictor, pertaining to a two-way association, is given by $\eta_{i,ts,kl} = \gamma + \gamma_{ts} + \gamma_{tk} + \gamma_{sl} + \delta_{tk}\delta_{sl}$, including an overall intercept, effects specific to times t and s : γ_{ts} , ‘row’ and ‘column’ effects γ_{tk} and γ_{sl} and multiplicative interactions. Obviously, this model is overparameterizing the association, making the use of restrictions necessary.

APPENDIX II: NEWTON-RAPHSON ALGORITHM

Replacing the matrix $W(\beta)$ in (21) by the matrix of second derivatives $H(\beta)$ of the log-likelihood (16) implements a Newton-Raphson algorithm. We present an expression for $H = H(\beta)$. From

McCullagh⁴⁰ it follows that the (p, q) element of H is

$$H_{pq} = \sum_{i=1}^N \sum_{t,r,s,h} \frac{\partial \eta_{it}}{\partial \beta_p} \frac{\partial \mu_{ir}}{\partial \eta_{it}} \frac{\partial^2 l}{\partial \mu_{ir} \mu_{is}} \frac{\partial \mu_{is}}{\partial \eta_{ih}} \frac{\partial \eta_{ih}}{\partial \beta_q} + \sum_{i=1}^N \sum_{t,h,r} \left[\frac{\partial \eta_{it}}{\partial \beta_p} \frac{\partial \eta_{ih}}{\partial \beta_q} \frac{\partial^2 \mu_{ir}}{\partial \eta_{it} \partial \eta_{ih}} + \frac{\partial^2 \eta_{it}}{\partial \beta_p \partial \beta_q} \frac{\partial \mu_{ir}}{\partial \eta_{it}} \right] \frac{\partial l}{\partial \mu_{ir}}.$$

Observing

$$\begin{aligned} \frac{\partial l}{\partial \mu_{ir}} &= \sum_k (V_i^{-1})_{rk} S_{ik} \\ \frac{\partial^2 l}{\partial \mu_{ir} \mu_{is}} &= -n_i * (V_i^{-1})_{rs} - \sum_{a,b,k} (V_i^{-1})_{ra} J_{i,s,ab} (V_i^{-1})_{bk} S_{ik} \\ J_{i,s,ab} &= \delta_{s,(a,b)} - \delta_{sa} \mu_{ib} - \delta_{sb} \mu_{ia} \end{aligned}$$

where δ is the Kronecker delta function and $l(t, h) = s$ if $\min(t_j, h_j) = s_j$ for all components of the index vectors, we can separate the terms involving S_i in the expression for $H(\beta)$:

$$H_{pq} = -W_{pq} + \sum_{i=1}^N \alpha_{ipq}^T S_i.$$

Obviously, the second term has expectation zero.

The first and second derivatives of μ_i with respect to v_i follow from the identities

$$\begin{aligned} \delta_{rs} &= \frac{\partial \mu_{ir}}{\partial \eta_{it}} \frac{\partial \eta_{it}}{\partial \mu_{is}} \\ \frac{\partial^2 \mu_{ir}}{\partial \eta_{it} \partial \eta_{ih}} &= - \sum_{s,t,v} \frac{\partial^2 \eta_{is}}{\partial \mu_{it} \partial \mu_{iv}} \frac{\partial \mu_{ir}}{\partial \eta_{is}} \frac{\partial \mu_{it}}{\partial \eta_{it}} \frac{\partial \mu_{iv}}{\partial \eta_{ih}}. \end{aligned}$$

Note that the first identity merely rephrases that $Q_i Q_i^{-1} = I$.

If we opt for linear predictors, we obtain

$$\frac{\partial \eta_i}{\partial \beta} = X_i \quad \frac{\partial^2 \eta_{it}}{\partial \beta_p \partial \beta_q} = 0.$$

We are now able to rewrite the Hessian in a concise matrix form

$$H(\beta) = \sum_{i=1}^n X_i^T [F_i + (Q_i^T)^{-1} G_i (Q_i)^{-1}] X_i$$

with

$$\begin{aligned} F_i &= \left(\sum_r \frac{\partial^2 \mu_{ir}}{\partial \eta_{it} \partial \eta_{ih}} \frac{\partial l}{\partial \mu_{ir}} \right)_{t,h} \\ G_i &= \frac{\partial^2 l}{\partial \mu_i \partial \mu_i^T}. \end{aligned}$$

Finally, if we again choose a link function of the type (6) we can use simple forms

$$Q_i = \frac{\partial \eta_i}{\partial \mu_i} = C \{ \text{diag}(A \mu_i) \}^{-1} A = C B_i A$$

and

$$\frac{\partial^2 \eta_{ij}}{\partial \mu_i \mu_i^T} = -A^T B_{ij}^{(2)} A$$

where the matrix $B_{ij}^{(2)}$ is obtained by multiplying all rows of B^2 with the j th row of C .

REFERENCES

1. Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. 'A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data', *International Statistical Review* **59**, 25–35 (1991).
2. Liang, K.-Y., Zeger, S. L. and Qaqish, B. 'Multivariate regression analyses for categorical data', *Journal of the Royal Statistical Society, Series B*, **54**, 3–40 (1992).
3. Ashford, J. R. and Sowden, R. R. 'Multivariate probit analysis', *Biometrics*, **26**, 535–546 (1970).
4. Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. and Lehnen, R. G. 'A general methodology for the analysis of experiments with repeated measurements of categorical data', *Biometrics*, **33**, 133–158 (1977).
5. Liang, K.-Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
6. Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
7. Robins, J., Rotnitzky, A. and Zhao, L. P. 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association*, **90**, 106–121 (1995).
8. Bahadur, R. R. 'A representation of the joint distribution of responses to n dichotomous items', in Solomon, H. (ed.), *Studies in Item Analysis and Prediction*, Stanford Mathematical Studies in the Social Sciences VI, Stanford University Press, Stanford, California, 1961.
9. Cox, D. R. 'The analysis of multivariate binary data', *Applied Statistics*, **21**, 113–120 (1972).
10. Prentice, R. L. 'Correlated binary regression with covariates specific to each binary observation', *Biometrics*, **44**, 1033–1048 (1988).
11. Zhao, L. P. and Prentice, R. L. 'Correlated binary regression using a quadratic exponential model', *Biometrika*, **77**, 642–648 (1990).
12. Zhao, L. P., Prentice, R. L. and Self, S. G. 'Multivariate mean parameter estimation by using a partly exponential model', *Journal of the Royal Statistical Society B*, **54**, 805–811 (1992).
13. Fitzmaurice, G. M. and Laird, N. M. 'A likelihood-based method for analysing longitudinal binary responses', *Biometrika*, **80**, 141–151 (1993).
14. Fitzmaurice, G. M., Laird, N. M. and Rotnitzky A. 'Regression models for discrete longitudinal responses', *Statistical Science*, **8**, 284–309 (1993).
15. Kauermann, G. 'Notes on multivariate logistic models for contingency tables', Technical Report, Forschungsberichte der Fachbereich Informatik, Technical University Berlin, 1993.
16. Barndorff-Nielsen, O. E. *Information and Exponential Families in Statistical Theory*, Wiley, Chichester, 1978.
17. Ekholm, A. 'Fitting regression models to a multivariate binary response', in Rosenqvist, G., Juselius, K., Nordström, K. and Palmgren, J. (eds), *A Spectrum of Statistical Thought: Essays in Statistical Theory, Economics, and Population Genetics in Honour of Johan Fellman*, Swedish School of Economics and Business Administration, Helsinki, 1991, pp. 19–32.
18. Dale, J. R. 'Global cross-ratio models for bivariate, discrete, ordered responses', *Biometrics*, **42**, 909–917 (1986).
19. Plackett, R. L. 'A class of bivariate distributions', *Journal of the American Statistical Association*, **60**, 516–522 (1965).
20. Lipsitz, S. R., Laird, N. M. and Harrington, D. P. 'Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association', *Biometrika*, **78**, 153–160 (1991).
21. Molenberghs, G. and Lesaffre, E. 'Marginal modelling of correlated ordinal data using a multivariate Plackett distribution', *Journal of the American Statistical Association*, **89**, 633–644 (1994).
22. McCullagh, P. and Nelder, J. A. *Generalized Linear Models*, Chapman & Hall, London, 1989.

23. Lang, J. B. and Agresti, A. 'Simultaneously modeling joint and marginal distributions of multivariate categorical responses', *Journal of the American Statistical Association*, **89**, 625–632 (1994).
24. Balagtas, C. C., Becker, M. P. and Lang, J. B. 'Marginal modeling of categorical data from cross-over experiments', *Applied Statistics*, **44**, 63–77 (1994).
25. Haber, M. and Brown, M. 'Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints', *Journal of the American Statistical Association*, **81**, 477–482 (1986).
26. Glonek, G. F. V. and McCullagh, P. 'Multivariate logistic models', *Journal of the Royal Statistical Society, Series B*, **57**, 533–546 (1995).
27. Deming, W. E. and Stephan, F. F. 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *Annals Mathematical Statistics*, **11**, 427–444 (1940).
28. Agresti, A. and Lang, J. B. 'A proportional odds model with subject-specific effects for repeated ordered categorical responses', *Biometrika*, **80**, 527–534 (1993).
29. Grizzle, J. E., Starmer, C. F. and Koch, G. G. 'Analysis of categorical data by linear models', *Biometrics*, **25**, 189–195, (1969).
30. Lesaffre, E. and Molenberghs, G. 'Multivariate probit analysis: a neglected procedure in medical statistics', *Statistics in Medicine*, **10**, 1391–1403 (1991).
31. Shervish, M. 'Multivariate normal probabilities with error bound', *Applied Statistics*, **33**, 81–94 (1984).
32. Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, 1975.
33. Dale, J. R. 'Local versus global association for bivariate ordered responses', *Biometrika*, **71**, 507–514 (1984).
34. Csiszar, I. 'I-divergence geometry of probability distributions and minimisation problems', *Annals of Probability*, **3**, 146–158 (1975).
35. Goodman, L. A. 'Association models and canonical correlation in the analysis of cross-classifications having ordered categories', *Journal of the American Statistical Association*, **74**, 537–552 (1981).
36. White, H. 'Maximum likelihood estimation of misspecified models', *Econometrica*, **50**, 1–25 (1982).
37. Kenward, M. G., Lesaffre, E. and Molenberghs, G. 'An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random', *Biometrics*, **50**, 945–953 (1994).
38. Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997) 'The analysis of longitudinal ordinal data with non-random dropout', *Biometrika*, **84**, 33–44 (1997).
39. Fitzmaurice, G. M. and Lipsitz, S. R. 'A model for binary time series data with serial odds ratio patterns', *Applied Statistics*, **44**, 51–61 (1995).
40. McCullagh, P. *Tensor Methods in Statistics*, Chapman & Hall, London, 1987.