A penalized approach for the bivariate ordered logistic model with applications to social and medical data.

Marco Enea [1 2], and Gianfranco Lovison [1]

[1] Dipartimento di Scienze Economiche Aziendali e Statistiche, University of Palermo, Palermo, Italy

[2] Istituto per l'Ambiente Marino Costiero, Consiglio Nazionale delle Ricerche, Mazara del Vallo, Italy

---

Address for correspondence: Marco Enea, Dipartimento di Scienze Economiche, Aziendali e Statistiche - Universitá degli Studi di Palermo, Viale delle Scienze, Edificio 13 - 91028 Palermo, Italy.

E-mail: marco.enea@unipa.it.

Phone: (+39) 09123895280.

Fax: .

Abstract: Bivariate ordered logistic models (BOLMs) are appealing to jointly model the marginal distribution of two ordered responses and their association, given a set of covariates. When the number of categories of the responses increases, the number of global odds ratios to be estimated also increases, and estimation gets problematic.

In this work we propose a nonparametric approach for the maximum likelihood estima-

tion of a BOLM wherein penalties to the differences between adjacent row and column effects are applied. Our proposal is then compared to the Goodman's model and the Dale's model. Some simulation results as well as analyses of two real data sets are presented and discussed.

Key words: Dale model; bivariate ordered logistic model; penalized maximum likelihood estimation; ordinal association

1 Introduction

Models for association play a central role in ordered categorical data analysis. For the multivariate case, marginal models (MMs) represent a natural choice to model marginal distributions of the responses given covariates. An example of full likelihood based marginal model is Dale (1986). A similar model, the multivariate logistic model described in Glonek and McCullagh (1995), but restricted to the bivariate ordered version, is the basis on which we develop our proposal. Some open, or at least not completely solved, problems about estimation of a multivariate ordered logistic model are of computational type and concern maximum likelihood (ML) estimation by iterative algorithms, often providing invalid estimates at the kth step, exceeding the boundaries of the parameter space. Some of such problems could be solved as in Colombi and Forcina (2001) and Bartolucci and Forcina (2002) by including strict inequality constraints. However, constrained ML estimation is appealing only when a particular application implies natural ordering constraints. On the contrary, when the ordering is not fully reliable, or externally imposed, like in responses which arise from discretized versions of latent continuous variables, using inequality con-

straints may not be appropriate. Indeed, due to lack of subject-matter knowledge that yields natural restrictions on marginal distributions, no strict ordering constraints are appropriate, and more helpful and flexible approaches are necessary. In these situations, a nonparametric approach may be useful (Dardanoni and Forcina, 1998). Within the possible range of nonparametric approaches, penalization is the one considered in this paper. Surprisingly, there is little literature on penalization applied to marginal models. Desantis et al. (2008) apply a ridge penalty to a latent class model for ordinal data to stabilize ML estimation, that would otherwise not be computationally feasible without application of strict constraints. Other contributions deal mainly with forms of longitudinal (Gieger, 1997; Fahrmeier et al., 1999) or horizontal (Bustami et al., 2001) nonparametric modelling. The former focuses on smoothing of variation of marginal and association parameters over time, the latter refers to a form of smoothing on covariates, often by using splines.

Our proposal is based on a form of vertical smoothing - that is across response levels - of the regression parameters in order to regularize the parameter space and/or fit polynomial models using scores "chosen by the data". After recalling the Dale, the Gloneck-McCullagh and the bivariate partial proportional odds models Section 2 introduces the penalized ML estimation approach and the penalty terms we propose. Section 3 deals with hypothesis testing and the asymptotic distribution of the penalized log likelihood ratio test. The theoretical results of Section 3 and the performance of the approach is shown by simulation Section 5. Two applications are considered in Section 6: in the first one, we compare our proposal to the Dale (1986) and the Goodman (1979) models on a literature data set on social mobility, whereas the second application is about the analysis of a data set of liver disease patients.

2 Bivariate ordered logit models

For two ordered outcomes A_1 and A_2, define the row and column marginal cumulative probabilities of a $D_1 \times D_2$ contingency table A_1A_2 as

$$\mu_{r.} = P(A_1 \leq r) = \sum_{i \leq r} \pi_{i.}, \quad \mu_{.c} = P(A_2 \leq c) = \sum_{j \leq c} \pi_{.j},$$

and the upper-left quadrant probabilities as

$$\mu_{rc} = P(A_1 \leq r, A_2 \leq c) \quad = \quad \sum_{i \leq r} \sum_{j \leq c} \pi_{ij},$$

with $r = 1, \ldots, D_1, c = 1, \ldots, D_2$. By differencing we obtain

$$\begin{aligned} P(A_1 \leq r, A_2 > c) &= \mu_{r.} - \mu_{rc}, \\ P(A_1 > r, A_2 \leq c) &= \mu_{.c} - \mu_{rc}, \\ P(A_1 > r, A_2 > c) &= 1 - \mu_{r.} - \mu_{.c} + \mu_{rc}. \end{aligned}$$

By choosing the cumulative odds as ordinal risk measures, and the logit as link function, we obtain the *global logits* (or *log global odds*):

$$\log \phi_{1r} = \text{logit}[P(A_1 \leq r)] = \log(\mu_{r.}) - \log(1 - \mu_{r.}), \tag{2.1}$$

$$\log \phi_{2c} = \text{logit}[P(A_2 \leq c)] = \log(\mu_{.c}) - \log(1 - \mu_{.c}), \tag{2.2}$$

$r = 1, \ldots, D_1 - 1, c = 1, \ldots, D_2 - 1$. By choosing the cross-products of quadrant probabilities as ordinal association measures, and the natural logarithm as link function, the *log global odds ratios* (or log-GORs) are defined as:

$$\begin{aligned} \log \psi_{rc} &= \log \frac{P(A_1 \leq r, A_2 \leq c) P(A_1 > r, A_2 > c)}{P(A_1 \leq r, A_2 > c) P(A_1 > r, A_2 \leq c)} \\ &= \log \frac{\mu_{rc}(1 - \mu_{r.} - \mu_{.c} + \mu_{rc})}{(\mu_{r.} - \mu_{rc})(\mu_{.c} - \mu_{rc})}. \end{aligned} \tag{2.3}$$

Given the three parameters $\mu_{r.}$, $\mu_{.c}$, and ψ_{rc}, we may find the corresponding joint cumulative probabilities with the following inversion formula:

$$\mu_{rc} = \begin{cases} \frac{1}{2}(\psi_{rc}-1)^{-1}(a_{rc} - \sqrt{a_{rc}^2 + b_{rc}}) & \text{if } \psi_{rc} \neq 1, \\ \mu_{r.}\mu_{.c} & \text{if } \psi_{rc} = 1, \end{cases} \tag{2.4}$$

where $a_{rc} = 1 + (\mu_{r.} + \mu_{.c})(\psi_{rc} - 1)$ and $b_{rc} = -4\psi_{rc}(\psi_{rc}-1)\mu_{r.}\mu_{.c}$. If the cumulative probabilities $\mu_{r.}$ and $\mu_{.c}$ satisfy the constraints $\mu_{r.} < \mu_{r+1,.}$ for $r = 1,\ldots,D_1-1$, and $\mu_{.c} < \mu_{.,c+1}$ for $c = 1,\ldots,D_2-1$, and the global odds ratios are not dependent on the category, that is $\psi_{rc} = \psi$, then (2.4) is a Plackett distribution (Plackett, 1965). Thus, the bivariate Dale regression model for $(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\psi}_{12})'$ is as follows:

$$\begin{cases} \log[\phi_{1,r}(\boldsymbol{x})] = \beta_{10r} - \boldsymbol{\beta}_1'\boldsymbol{x}, \\ \log[\phi_{2,c}(\boldsymbol{x})] = \beta_{20c} - \boldsymbol{\beta}_2'\boldsymbol{x}, \\ \log[\psi_{rc}(\boldsymbol{x})] = \alpha + \rho_{1r} + \rho_{2c} + \sigma_{rc} - \boldsymbol{\beta}_3'\boldsymbol{x}, \end{cases} \tag{2.5}$$

$r = 1,...,D_1-1$, $c = 1,...,D_2-1$. This model does not require marginal scores for responses and it is also invariant under any monotonic transformation of the marginal responses. Further, since the model is based on global odds ratios, collapsing adjacent row or column categories does not produce any effect in parameter interpretation, which remains unchanged with the exception of the intercepts related to the collapsed categories. This is in contrast with the RC Goodman model which uses local cross-ratios. In a more general framework than (2.5), Glonek and McCullagh (1995) introduce the *multivariate logistic model*:

$$\boldsymbol{C}'\log(\boldsymbol{L}\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{2.6}$$

where $\boldsymbol{C}$ is a contrasts matrix, $\boldsymbol{L}$ is a matrix with elements $a_{ij} \geq 0$ such that $\boldsymbol{L}\boldsymbol{\pi} = \boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{C}'\log(\boldsymbol{L}\boldsymbol{\pi})$ is the parameter vector of interest, and $\boldsymbol{X}$, an $n \times p$ matrix, with $n =$

$\prod_{k=1}^{K} D_k$. Although formulation (2.6) is referred to $K \geq 2$ responses, here only two responses A_1 and A_2 are considered. The components of $\boldsymbol{C}' \log(\boldsymbol{L}\boldsymbol{\pi})$ are symbolically denoted by $\boldsymbol{\eta} = (\eta_\varnothing, \boldsymbol{\eta}'_{A_1}, \boldsymbol{\eta}'_{A_2}, \boldsymbol{\eta}'_{A_1A_2})'$, where $\eta_\varnothing = \log(\sum \boldsymbol{\pi}) = 0$ is the null contrast and the remaining vectors have elements specified by (2.1), (2.2) and (2.3), respectively. We will refer to (2.6) as the bivariate ordered logistic model (BOLM). Lapp et al. (1998) show how to fit the Dale and Goodman models starting from the framework of a BOLM. Some computational problems may arise when fitting a multivariate logistic model, depending on the number of responses and categories. For example, when inverting equation $\boldsymbol{\eta} = \boldsymbol{C}' \log(\boldsymbol{L}\boldsymbol{\pi})$ to obtain $\boldsymbol{\pi}$ in terms of $\boldsymbol{\eta}$, it may happen that for certain fixed values of $\boldsymbol{\eta}$ no positive solution $\boldsymbol{\pi}$ exists. Although $\boldsymbol{\pi} > 0$ ensures the matrix $\boldsymbol{C}'\boldsymbol{D}^{-1}\boldsymbol{L}$ to be invertible (Glonek and McCullagh, 1995, Theorem 1), where $\boldsymbol{D} = \text{diag}(\boldsymbol{L}\boldsymbol{\pi})$, the range of the mapping is not a hyper-rectangle and fixing some components of $\boldsymbol{\eta}$ restricts the range of the remaining components, i.e. the model is not *variation independent*. Although this problem is particularly magnified for $K > 2$ responses (Bergsma and Rudas, 2002; Qaqish and Ivanova, 2006), computational problems can also arise in the bivariate case, above all when considering certain particular model configurations. For instance, it may happen that not only the intercepts but some covariates have a category-dependent effect. To highlight this effect one may want to fit a bivariate version of the partial proportional odds model proposed by Peterson and Harrell (1990). However, such a model can be computationally very hard to fit, even with a limited and reasonable number of parameters. To deal with this difficulty, we propose to regularize the parameter space by penalizing the log-likelihood of the model. This allows to increase the range of possible models to be fitted. The penalty term we use for this is introduced in Section 2.1.

The fit of a BOLM becomes computationally hard also when the number of response categories increases. In addition, the model may result overparameterized. Lapp et al. (1998)

fit a Dale's model by imposing constraints on the row and column interactions of the association intercepts in order to reduce the number of parameters. However, this type of data appears to be too "rich" to be modeled with fully parametric models and nonparametric or semiparametric models, followed by graphical presentation, could result more useful (Eilers and Marx, 1996). In order to smooth the marginal and association effects across the response categories, in Section 2.2, a penalty term for nonparametric modelling is introduced. Such term, often employed in the P-spline context (Eilers et al., 2006), has been suitably re-written to be used in the framework of a BOLM. In part, this approach can be considered the bivariate extension of the models proposed by Tutz (2003).

The ordinal nature of the responses imposes inequality constraints on marginal distributions which have to be taken into account in model estimation. In Section 2.3, we present a penalty term, which is able to mimic such inequality constraints.

In order to better understand the potential of the penalization approach, some further notation is needed, according to that used in Tutz and Scholz (2003) for the univariate cumulative logistic regression model. Let $\mathscr{Q}$ be the set of indices of all the covariates, excluding the intercepts, and $\mathscr{P} \subset \mathscr{Q}$ be a subset of p covariates. Let $\mathscr{S}$ be the set of indices of the variables whose effects we assume do not depend on categories and such that $\mathscr{S} \subseteq \mathscr{P}$, and let $\bar{\mathscr{S}} = \mathscr{P} \backslash \mathscr{S}$. In particular, we define $\mathscr{S}$ and $\bar{\mathscr{S}}$ as $\mathscr{S} = \cup_{k=1}^{3} \mathscr{S}_k$, and $\bar{\mathscr{S}} = \cup_{k=1}^{3} \bar{\mathscr{S}}_k$, where $\mathscr{S}_k$ and $\bar{\mathscr{S}}_k$ are the subsets of $\mathscr{S}$ and $\bar{\mathscr{S}}$, respectively, associated to the kth equation. To complete the notation, let $\mathscr{S}^0 = \{0\} \cup \mathscr{S}$ and $\bar{\mathscr{S}}^0 = \{0\} \cup \bar{\mathscr{S}}$. Consider the following model where only a part of the covariates is supposed to be category-independent:

$$\begin{cases} \log[\phi_{1r}(\boldsymbol{x}_i)] = \beta_{10r} + \boldsymbol{\beta}'_{1,\mathscr{S}_1}\boldsymbol{x}_{i,\mathscr{S}_1} + \boldsymbol{\beta}'_{1,\bar{\mathscr{S}}_1 r}\boldsymbol{x}_{i,\bar{\mathscr{S}}_1}, \\ \log[\phi_{2c}(\boldsymbol{x}_i)] = \beta_{20c} + \boldsymbol{\beta}'_{2,\mathscr{S}_2}\boldsymbol{x}_{i,\mathscr{S}_2} + \boldsymbol{\beta}'_{2,\bar{\mathscr{S}}_2 c}\boldsymbol{x}_{i,\bar{\mathscr{S}}_2}, \\ \log[\psi_{rc}(\boldsymbol{x}_i)] = \beta_{30rc} + \boldsymbol{\beta}'_{3,\mathscr{S}_3}\boldsymbol{x}_{i,\mathscr{S}_3} + \boldsymbol{\beta}'_{3,\bar{\mathscr{S}}_3 rc}\boldsymbol{x}_{i,\bar{\mathscr{S}}_3}, \end{cases} \tag{2.7}$$

$(r = 1,\dots,D_1-1,\ c = 1,\dots,D_2-1)$. We refer to model (2.7) as the *Non-Uniform association and Partially Proportional Odds Model* (NUPPOM). Although in the univariate case the phrase "proportional odds" is usually referred to a model with covariate effects which do not depend on the categories, here we will refer to a *Uniform association and Proportional Odds Model* (UPOM) as a model defined from (2.7) assuming $\beta_{30rc} = \beta_{30}$ and $\bar{\mathscr{S}} = \varnothing$. On the other hand, a *Non-Uniform association and Non-Proportional Odds Model* (NUNPOM) will be defined from (2.7) assuming $\mathscr{S} = \varnothing$ and with category-dependent association intercepts. Note that the intercepts for the marginal equations (that is the global-logit intercepts) are never supposed to be independent of the categories, whatever the model. According to these definitions the bivariate Dale model (2.5) is a NUPOM, and it becomes a UPOM when $\rho_{1r} = 0$, $\rho_{2c} = 0$ and $\sigma_{rc} = 0$, $r = 1,\dots,D_1-1,\ c = 1,\dots,D_2-1$. Further, to specify that a NUPPOM is fitted we will also write $NUPPOM(\mathscr{S})$ and to indicate that a UPOM is fitted we will also write $UPOM(\bar{\mathscr{S}}^0)$.

Under multinomial sampling with frequencies $\boldsymbol{y}_i \sim M(n_i, \boldsymbol{\pi}_i)$, consider the model $\boldsymbol{C}'\log(\boldsymbol{L}\boldsymbol{\pi}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, with the matrices $\boldsymbol{C}$ and $\boldsymbol{L}$ such that the marginal parameters are *global logits*, the association parameters are *log global odds ratios*, and the constraint $\sum_{j=1}^{D_1}\sum_{k=1}^{D_2}\pi_{ijk} = 1$ is included. Then, the kernel of the log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{m} l(\boldsymbol{\beta};\boldsymbol{y}_i) = \sum_{i=1}^{m} \boldsymbol{y}_i'\log(\boldsymbol{\pi}_i), \tag{2.8}$$

where m, the observed number of response configurations, is such that $\sum_{i=1}^{m} n_i = n$, with n indicating the sample size. The *penalized log-likelihood* has the form

$$l_P(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\tau(\boldsymbol{\beta}), \tag{2.9}$$

where $\tau(\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta}$, and $\mathbf{P}$ represents the penalization and includes the smoothing parameter. Penalized ML estimation formulas are given and discussed in Appendix A, while Appendix B shows the matrix form of $\mathbf{P}$.

2.1 Penalty terms for parameter space regularization

When the cross-tabulation of the responses contains one or more zeros, parameter estimation by Fisher-scoring may be challenging at each iteration. In these cases, one may try to reduce l_{step}, the step length (see Appendix A). However, estimates of the association structure may result to be too irregular, with very high (or very low) estimated odds ratios in correspondence of the cell with zeros. In order to stabilize the ML estimates of the BOLM using Fisher scoring, a reduction of the parameter space may be helpful. We propose to penalize both the marginal and the association parameters. In addition, since the model is not variation independent, applying a penalty term on association parameters might be useful to limit the range of the possible values that the marginal parameters can assume, so avoiding a failure of the Fisher scoring. The general expression of $\tau(\boldsymbol{\beta})$ is:

$$\begin{aligned}\tau(\boldsymbol{\beta}) &= \sum_{j\in\mathscr{T}_1^0}\lambda_{1j}\sum_{r=r_\zeta}^{D_1-1}\zeta(\beta_{1jr}) + \sum_{j\in\mathscr{T}_2^0}\lambda_{2j}\sum_{c=c_\zeta}^{D_2-1}\zeta(\beta_{2jc}) \\ &+ \sum_{j\in\mathscr{T}_3^0}\lambda_{3j}\sum_{r=r_\zeta}^{D_1-1}\sum_{c=c_\zeta}^{D_2-1}\zeta(\beta_{3jrc}),\end{aligned} \tag{2.10}$$

where λ_{kj} is the smoothing parameter for the jth variable of the kth equation of system (2.7), $k = 1,2,3$, and $\zeta(\cdot)$ is a generic function that characterizes the penalty term. Notice

that the starting values r_ζ and c_ζ of the summation indices depend on $\zeta(\cdot)$. In the following subsections, we introduce three different specifications of $\zeta(\cdot)$.

2.1.1 The ARC1 penalty term

A first specification of $\zeta(\cdot)$ is $\zeta(\alpha_t) = (\Delta\alpha_t)^2$, where Δ is the order 1 difference operator, that is $\Delta\alpha_t = \alpha_t - \alpha_{t-1}, t \geq 2$. This penalty term involves the differences of Adjacent Row and Column parameters (ARC1). The term is defined as

$$\begin{aligned} \tau(\boldsymbol{\beta}) &= \sum_{j\in\bar{\mathscr{I}}_1^0} \lambda_{1j} \sum_{r=2}^{D_1-1} (\Delta\beta_{1jr})^2 + \sum_{j\in\bar{\mathscr{I}}_2^0} \lambda_{2j} \sum_{c=2}^{D_2-1} (\Delta\beta_{2jc})^2 \\ &+ \sum_{j\in\bar{\mathscr{I}}_3^0} \lambda_{3j} \sum_{r=2}^{D_1-1} \sum_{c=2}^{D_2-1} (\Delta\beta_{3jrc})^2, \end{aligned} \tag{2.11}$$

and it is aimed at overcoming estimation problems by reducing parameter space. As $\lambda_{kj} \to \infty$, $k = 1,2,3$, $\forall j \in \bar{\mathscr{I}}_k^0$, all the parameters indexed by j will tend to be equal among the categories. In practice, if all category dependent parameters are involved in the penalization, the model will tend to a *UPOM* for high smoothing values. Although (2.11) allows to penalize marginal intercepts, it is preferable to avoid a strong penalization on such parameters, in order not to violate (2.15).

The choice of λ is two steps: the first step is based on the minimum value λ_{min} for which Fisher scoring does not fail. The simulation study in Section 5 will clarify this choice. In the second step, the search of an optimal λ, say λ_{opt}, satisfying criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), is then performed on the interval $[\lambda_{min}, +\infty)$.

2.1.2 Ridge-type penalty

Another specification of $\zeta(\cdot)$ aimed at reducing the parameter space is $\zeta(\alpha) = \alpha^2$, corresponds to a ridge-type penalty for the bivariate logistic regression model:

$$\begin{aligned}\tau(\boldsymbol{\beta}) &= \sum_{j\in\bar{\mathscr{I}}_1^0}\lambda_{1j}\sum_{r=1}^{D_1-1}\beta_{1jr}^2+\sum_{j\in\bar{\mathscr{I}}_2^0}\lambda_{2j}\sum_{c=1}^{D_2-1}\beta_{2jc}^2 \\ &+ \sum_{j\in\bar{\mathscr{I}}_3^0}\lambda_{3j}\sum_{r=1}^{D_1-1}\sum_{c=1}^{D_2-1}\beta_{3jrc}^2. \end{aligned} \tag{2.12}$$

For $\lambda_{kj}\to\infty$, $k=1,2,3$, $\forall j\in\bar{\mathscr{I}}_k^0$, all parameters indexed by j will tend to zero. A similar penalty, not involving the third term, is used by Desantis et al. (2008) in a penalized latent class model for ordinal data.

2.1.3 Lasso-type penalty

In this paper, emphasis for regularization problems is on ARC1, but many other penalty terms are possible. For example, a "horizontal" lasso-type penalization is written as

$$\begin{aligned}\tau(\boldsymbol{\beta}) &= \sum_{j\in\bar{\mathscr{I}}_1^0}\lambda_{1j}\sum_{r=1}^{D_1-1}|\beta_{1jr}|+\sum_{j\in\bar{\mathscr{I}}_2^0}\lambda_{2j}\sum_{c=1}^{D_2-1}|\beta_{2jc}| \\ &+ \sum_{j\in\bar{\mathscr{I}}_3^0}\lambda_{3j}\sum_{r=1}^{D_1-1}\sum_{c=1}^{D_2-1}|\beta_{3jrc}|. \end{aligned} \tag{2.13}$$

For $\lambda_{kj}\to\infty$, $k=1,2,3$, $\forall j\in\bar{\mathscr{I}}_k^0$, all the parameters indexed by j will tend to zero. The lasso penalty may be used as an alternative to (2.12).

A further specification of $\zeta(\cdot)$ could be $\zeta(\alpha_t)=|\Delta\alpha_t|$, and $r_\zeta = c_\zeta = 2$, a "vertical" lasso-type penalty on the differences of adjacent parameters, as an alternative to (2.11).

2.2 A penalty term for nonparametric modelling

Beside being useful for reducing the parameter space and for reproducing a $UPOM$, the following generalization of the penalty term ARC1, hereafter denoted by ARC2, can be used to specify row or column effects and to fit nonparametric models where the effects are determined by a polynomial:

$$\begin{aligned}\tau(\boldsymbol{\beta}) &= \sum_{j\in\bar{\mathscr{S}}_1^0}\lambda_{1j}\sum_{r=s_1+1}^{D_1-1}(\Delta^{s_{1j}}\beta_{1jr})^2+\sum_{j\in\bar{\mathscr{S}}_2^0}\lambda_{2j}\sum_{c=s_2+1}^{D_2-1}(\Delta^{s_{2j}}\beta_{2jc})^2\\ &+\sum_{j\in\bar{\mathscr{S}}_3^0}\left[\lambda_{3j}\sum_{r=s_3+1}^{D_1-1}\sum_{c=1}^{D_2-1}(\Delta^{s_{3j}}\beta_{3jrc})^2\right.\\ &+\left.\lambda_{4j}\sum_{r=1}^{D_1-1}\sum_{c=s_4+1}^{D_2-1}(\Delta^{s_{4j}}\beta_{3jrc})^2\right],\end{aligned} \tag{2.14}$$

where $\Delta^a=\Delta(\Delta^{a-1})$. Consider the following penalty settings:

- as $\lambda_{hj}=0$, $h=1,...,4$, $\forall j\in\bar{\mathscr{S}}^0$, an unrestricted model will be fitted;
- as $\lambda_{hj}\to\infty$, $h=1,...,4$, $\forall j\in\bar{\mathscr{S}}_1\cup\bar{\mathscr{S}}_2\cup\bar{\mathscr{S}}_3^0$ and $s_{hj}=1$, the fitted parameters will tend to be equal, and the model will tend to a $UPOM$;
- as $\lambda_{3j}\to\infty$, $\lambda_{4j}=0$, $\forall j\in\bar{\mathscr{S}}_3^0$ and $s_{3j}=1$, a model with column effects will be fitted;
- as $\lambda_{3j}=0$, $\lambda_{4j}\to\infty$, $\forall j\in\bar{\mathscr{S}}_3^0$ and $s_{4j}=1$, a model with row effects will be fitted;
- as $\lambda_{hj}\to\infty$, $h=1,2$, $\forall j\in\bar{\mathscr{S}}^0$ and $s_{hj}>1$, the fitted parameters will follow a polynomial curve of degree $s_{hj}-1$.
- as $\lambda_{hj}\to\infty$, $h=3,4$, $\forall j\in\bar{\mathscr{S}}^0$ and $s_{hj}>1$, the fitted parameters will follow a polynomial surface of degree $s_{3j}+s_{4j}-2$.

Notice the difference between the penalty terms included in (2.14) and those included in the penalized log-likelihood (14) in Tutz (2003), suggested for a single ordered response. In that paper, the author proposed to penalize the differences of adjacent categories, for a vertical smoothing, jointly to the use of penalized B-splines for a horizontal smoothing, resulting in a form similar to (2.14). Also notice the differences with the bivariate horizontal smoothing approach by Bustami et al. (2001) which presented the additive bivariate Dale model, for continuous, category-independent covariates, as a natural extension of the generalized additive model (Hastie and Tibshirani, 1990).

Penalty (2.14) may be useful to assume certain dependence structures on the categories, for both marginal and association parameters. For example, if one wants to assume a linear trend for the row marginal effects, one may assume $\eta_{1ir} = \beta_{10r} + \sum_{j=1}^{p} x_{ij}\beta_{1jr}$, where $\beta_{1jr} = \alpha_{0j} + \alpha_{1j}\delta_{jr}$, with α_{0j} and α_{1j} unknown parameters, and with scores δ_{jr}. In spite of its simplicity, such an approach assumes arbitrary scores. An alternative way is just to use a penalization approach with penalty term ARC2 which uses scores "chosen by the data" (Tutz and Scholz, 2003). Indeed, the smoothing parameters and the polynomial degrees can be chosen on the basis of some criterion, such as the values that minimize the AIC. As a special case, suppose to want to fit a model which assumes a linear trend for the marginal parameters and an association structure composed by the interaction of two first degree polynomials[1]. By assuming, for simplicity, that the same variable x_j is present in all the equations of system (2.7), choosing $s_{hj} = 2, h = 1,\ldots,4$ and high smoothing values, for

[1]The degree of a two-variable polynomial is defined as the highest degree of its terms, and the degree of a term is the sum of the exponents of the variables that appear in it. Since (2.14) allows to fit only polynomial models with interactions, to distinguish each of the possible models having the same degree, it is more practical for us to indicate a model by specifying both the degrees of the one-variable polynomials, omitting to specify the (implicit) presence of interaction terms.

instance 10^8, the predictor becomes

- $x_{ij}\beta_{jr} = x_{ij}\gamma_{01j} + x_{ij}(\gamma_{11j}\cdot r)$
- $x_{ij}\beta_{jc} = x_j\gamma_{02j} + x_{ij}(\gamma_{12j}\cdot c)$
- $x_{ij}\beta_{jrc} = x_{ij}\gamma_{03j} + x_j(\gamma_{13j}\cdot r) + x_{ij}(\gamma_{23j}\cdot c) + x_{ij}(\gamma_{33j}\cdot r\cdot c)$,

with scores $\delta_{jr} = r$, $\delta_{jc} = c$, and $\boldsymbol{\delta}_{jrc} = (r,c)'$, that is pre-assigned equally-spaced scores.

2.3 Mimicking inequality constraints

The ordinal nature of the responses introduces some explicit ordering constraints on marginal distribution which have to be taken into account to avoid ill-conditioning of the predictor space. In particular, for the ith individual, such constraints are on the marginal predictors, that is

$$\beta_{k01} + \boldsymbol{\beta}'_{k1}\boldsymbol{x}_i < \beta_{k02} + \boldsymbol{\beta}'_{k2}\boldsymbol{x}_i < \ldots < \beta_{k0,D_k-1} + \boldsymbol{\beta}'_{k,D_k-1}\boldsymbol{x}_i, \tag{2.15}$$

$k = 1,2$. Although Lagrangians can be used to take into account such constraints, in the spirit of this paper, a penalized-oriented solution could be the following:

$$\tau(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[\sum_{k=1}^{2}\lambda_k\sum_{r=2}^{D_k-1} I(\Delta\eta_{kir})(\Delta\eta_{kir})^2\right], \tag{2.16}$$

where $\eta_{kir} = \beta_{k0r} + \boldsymbol{\beta}'_{kr}\boldsymbol{x}_i$, $\Delta\eta_{kir} = \eta_{kir} - \eta_{ki,r-1}$, and $I(z) = 1$ if $z \geq 0$, otherwise $I(z) = 0$. As $\lambda_k \to \infty$, the penalty term (2.16) acts in such a way to satisfy (2.15). The univariate version of (2.16) is used, for example, by Muggeo and Ferrara (2008) in a penalized splines context applied to univariate generalized linear models. It can also be used jointly to (2.10) or (2.14). Notice that, although seemingly superfluous, the inclusion of $(\Delta\eta_{kir})^2$ in (2.16)

derives from the necessity of writing $\tau(\boldsymbol{\beta})$ as a quadratic form in order to exploit the penalized ML formulae in Appendix A.

3 Hypothesis testing

When estimates are penalized, the asymptotic distribution of the penalized likelihood ratio (LR_P) statistic is known only for some hypothesis systems. As far as we know, neither exact nor asymptotic results are known for LR_P to check the hypothesis $(P)POM$, i.e. that of category-independent effects. We provide the conditions for which it is possible to approximate, under the hypothesis $(P)POM$, the LR_P asymptotic distribution by a χ^2 distribution. The assessment of this result is done through simulation studies carried out in Section 4. As an introduction, the next section reports a result already present in the literature, useful for a simple hypothesis system. To simplify notation we assume, without loss of generality, that the same index j refers to the same variable for both marginal and association equations.

3.1 The LR_P statistic for the hypothesis of null effects

Let us consider the specific partition of parameters $\boldsymbol{\beta}_{\mathscr{P}^0} = (\boldsymbol{\gamma}, \boldsymbol{\delta})'$, such that the null hypothesis:

$$H_0 : \boldsymbol{\delta} = \mathbf{0}, \tag{3.1}$$

postulates that only a subset of parameters is constrained. Furthermore, consider the penalized log-likelihood of the more general model, $l_P(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$, that of the reduced model, $l_P(\tilde{\boldsymbol{\gamma}}, \mathbf{0})$ and the penalized log-likelihood ratio statistic:

$$LR_P = -2\{l_P(\tilde{\boldsymbol{\gamma}}, \mathbf{0}) - l_P(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})\}. \tag{3.2}$$

Let $\mathbf{F}$ be the information matrix from the unpenalized partial likelihood, with subscripts denoting the submatrices, such as $\mathbf{F}_{\boldsymbol{\delta\delta}}$ for derivatives with respect to $\boldsymbol{\delta}$. Consider the matrix $\mathbf{F}_{\boldsymbol{\delta\delta}|\boldsymbol{\gamma}} = \mathbf{F}_{\boldsymbol{\delta\delta}} - \mathbf{F}_{\boldsymbol{\delta\gamma}}\mathbf{F}_{\boldsymbol{\gamma\gamma}}^{-1}\mathbf{F}_{\boldsymbol{\gamma\delta}}$. Then, under the null hypothesis, Gray (1994) shows the statistic LR_P has the same asymptotic distribution as $\sum \alpha_j Z_j^2$, where the Z_j's are independent standard Normal random variables, and the α_j's are the eigenvalues of the matrix $\lim_{n\to\infty} \mathbf{F}_{\boldsymbol{\delta\delta}|\boldsymbol{\gamma}}(\mathbf{F}_{\boldsymbol{\delta\delta}|\boldsymbol{\gamma}} + \mathbf{P})^{-1}$, where $\mathbf{P}$ is the matrix representing the penalty term. This approach has emerged to work satisfactorily in practice (Muggeo and Tagliavia, 2010).

3.2 The LR_P statistic for the $(P)POM$ hypothesis

Consider a full model of the *NUNPOM* type, i.e. for which all variables j, $j \in \mathscr{P}^0 \equiv \{\bar{\mathscr{P}}^0, \mathscr{S} = \varnothing\}$, have category-dependent effects, and a reduced model for which the effects of some variables j, $j \in \mathscr{S} \neq \varnothing$, are category independent. The penalized log-likelihood ratio test to check the hypothesis for comparing these two models, i.e. for testing the null hypothesis:

$$H_0 : \boldsymbol{\beta}_j = \beta_j \mathbf{1},\ j \in \mathscr{S}, \tag{3.3}$$

compares the maximum penalized log-likelihood $l_P(\hat{\boldsymbol{\beta}}_{\mathscr{P}^0})$, and the maximum penalized log-likelihood $l_P(\tilde{\boldsymbol{\beta}}_{\mathscr{S}}, \tilde{\boldsymbol{\beta}}_{\bar{\mathscr{P}}^0})$:

$$\begin{aligned} LR_P &= -2\{l_P(\tilde{\boldsymbol{\beta}}_{\mathscr{S}}, \tilde{\boldsymbol{\beta}}_{\bar{\mathscr{P}}^0}) - l_P(\hat{\boldsymbol{\beta}}_{\mathscr{P}^0})\} \\ &= 2\sum_{i=1}^{m}\sum_{r=1}^{D_1}\sum_{c=1}^{D_2} \mathbf{y}'_{irc} \log\left(\frac{\hat{\boldsymbol{\pi}}_{irc}}{\tilde{\boldsymbol{\pi}}_{irc}}\right) + \tau(\tilde{\boldsymbol{\beta}}) - \tau(\hat{\boldsymbol{\beta}}), \end{aligned} \tag{3.4}$$

where $\hat{\boldsymbol{\pi}}'_i = (\hat{\pi}_{i11}, \ldots, \hat{\pi}_{iD_1D_2})'$ is the estimated (by penalization) probability vector for the model under H_1 and $\tilde{\boldsymbol{\pi}}_i = (\tilde{\pi}_{i11}, \ldots, \tilde{\pi}_{iD_1D_2})'$ is the corresponding estimated (by penalization) probability vector for the reduced model. By supposing to use the penalty term ARC1

and by following Tutz and Scholz (2003), let $\lambda_{kj\mathscr{R}}\,(\lambda_{kj\mathscr{F}})$ denote the smoothing parameters for the reduced model (full model). Then, we have

$$\tau(\tilde{\boldsymbol{\beta}}) - \tau(\hat{\boldsymbol{\beta}}) =$$

$$\begin{aligned}
&= \sum_{j\in\bar{\mathscr{S}}^0}\Bigg[\lambda_{1j\mathscr{R}}\sum_{r=2}^{D_1-1}(\Delta\tilde{\beta}_{1jr})^2+\lambda_{2j\mathscr{R}}\sum_{c=2}^{D_2-1}(\Delta\tilde{\beta}_{2jc})^2\\
&+ \lambda_{3j\mathscr{R}}\sum_{r=2}^{D_1-1}\sum_{c=2}^{D_2-1}(\Delta\tilde{\beta}_{3jrc})^2\Bigg]-\sum_{j\in\mathscr{P}^0}\Bigg[\lambda_{1j\mathscr{F}}\sum_{r=2}^{D_1-1}(\Delta\hat{\beta}_{1jr})^2\\
&+ \lambda_{2j\mathscr{F}}\sum_{c=2}^{D_2-1}(\Delta\hat{\beta}_{2jc})^2+\lambda_{3j\mathscr{F}}\sum_{r=2}^{D_1-1}\sum_{c=2}^{D_2-1}(\Delta\hat{\beta}_{3jrc})^2\Bigg]\\
&= \sum_{j\in\bar{\mathscr{S}}^0}\Bigg\{\sum_{r=2}^{D_1-1}\left[\lambda_{1j\mathscr{R}}(\Delta\tilde{\beta}_{1jr})^2-\lambda_{1j\mathscr{F}}(\Delta\hat{\beta}_{1jr})^2\right]\\
&+ \sum_{c=2}^{D_2-1}\left[\lambda_{2j\mathscr{R}}(\Delta\tilde{\beta}_{2jc})^2-\lambda_{2j\mathscr{F}}(\Delta\hat{\beta}_{2jc})^2\right]\\
&+ \sum_{r=2}^{D_1-1}\sum_{c=2}^{D_2-1}\left[\lambda_{3j\mathscr{R}}(\Delta\tilde{\beta}_{3jrc})^2-\lambda_{3j\mathscr{F}}(\Delta\hat{\beta}_{3jrc})^2\right]\Bigg\}\\
&- \sum_{j\in\mathscr{S}}\Bigg[\lambda_{1j\mathscr{F}}\sum_{r=2}^{D_1-1}(\Delta\hat{\beta}_{kjr})^2+\lambda_{2j\mathscr{F}}\sum_{c=2}^{D_2-1}(\Delta\hat{\beta}_{kjc})^2\\
&+ \lambda_{3j\mathscr{F}}\sum_{r=2}^{D_1-1}\sum_{c=2}^{D_2-1}(\Delta\hat{\beta}_{kjrc})^2\Bigg].
\end{aligned}$$

If estimates are not penalized, that is if for $k=1,2,3$, $\lambda_{k1\mathscr{R}}=\lambda_{k2\mathscr{R}}=\cdots=\lambda_{k|\bar{\mathscr{S}}^0|\mathscr{R}}=\lambda_{k0\mathscr{F}}=\lambda_{k1\mathscr{F}}=\cdots=\lambda_{k|\mathscr{P}^0|\mathscr{F}}=0$, one obtains $\tau(\tilde{\boldsymbol{\beta}})-\tau(\hat{\boldsymbol{\beta}})=0$, and the LR_P statistic has the usual asymptotic χ^2 distribution. If $\lambda_{kj\mathscr{R}}=\lambda_{kj\mathscr{F}}$ is chosen for $j\in\bar{\mathscr{S}}^0$, $k=1,2,3$, then the first term is very small since $\tilde{\beta}_{kjr}\approx\hat{\beta}_{kjr}$, $\tilde{\beta}_{kjc}\approx\hat{\beta}_{kjc}$ and $\tilde{\beta}_{kjrc}\approx\hat{\beta}_{kjrc}$ for $r=1,\ldots,D_1-1, c=1,\ldots,D_2-1$. Thus, the fundamental term concerns the variables for which $j\in\mathscr{S}$ and, if estimates are penalized with a low smoothing value, converging to zero at an appropriate rate, the same asymptotic behaviour holds.

4 Simulation studies

The LR_P sampling distributions discussed in the previous sections are shown here by simulation. We set two simulation schemes, called *sim*1 and *sim*2, according to the two hypothesis systems in Sections 3.1 and 3.2, respectively. We generated $m = 1000$ pseudo-samples from a multinomial distribution with probability matrix $\Pi(\mathbf{X}\boldsymbol{\beta})$ and fitted reduced and unreduced models by penalizing the association parameter vector $\boldsymbol{\beta}_{32}$ in *sim*1, and the vector $\boldsymbol{\beta}_{30}$ of association intercepts in *sim*2. The simulation setting is:

- an equal number of levels in both responses and with values 3, 5, and 7;
- sample sizes n=200, 500, and 1000 when response levels are 3 or 5;
- sample sizes n=1000, 2000, and 5000 when response levels are 7;
- λ=0, 0.2, 0.5, 1, 2, 5, 10, 20, and 50;
- two binary covariates, X_1 and X_2, both sampled from a $Ber(0.5)$.

The set of λ values used is typical of *grid search* algorithms. We preferred not to include further covariates because, especially for the cases with either 5 or 7 levels for both responses, the computational complexity of the estimation exponentially increases for each factor level added. The criterion used to set the order of magnitude of n for the underlying table is to have, on average, 5 observations per cell. Accordingly, under a NUNPOM setting, for a simulation scheme with 3 levels per response and two binary variables, $n = 5 \times 3^2 \times 2^2 = 180$ (but it was rounded to 200). For the case with 5 levels $n = 5^3 \times 2^2 = 500$, while $n = 5 \times 7^2 \times 2^2 = 980$ (rounded to 1000) for the 7-level case.

4.1 Simulation scheme 1

In the first simulation scheme (*sim*1), we simulate the sampling distribution of the LR_P statistic under the null hypothesis of Section 3.1. The parameters chosen to generate the simulation depend on the number of response levels. In particular, these are specified as follows:

*Simulation sim*1*: case* $D_1 = D_2 = 3$

- $\boldsymbol{\beta}_{10} = (-0.5, 0.5)'$, $\boldsymbol{\beta}_{11} = (0.2, 0.2)'$, $\boldsymbol{\beta}_{12} = (-0.3, 0.3)'$,
- $\boldsymbol{\beta}_{20} = (-0.1, 0.6)'$, $\boldsymbol{\beta}_{21} = (0.3, 0.3)'$, $\boldsymbol{\beta}_{22} = (-0.2, 0.4)'$,
- $\boldsymbol{\beta}_{30} = (1.5, 2, 2.5, 3)'$, $\boldsymbol{\beta}_{31} = (-.5, -.5, -.5, -.5)'$, $\boldsymbol{\beta}_{32} = (0, 0, 0, 0)'$.

With this set of parameters the model under the null hypothesis $\beta_{32rc} = 0$, $r = 1, 2$, $c = 1, 2$ is a NUPPOM. The model under the alternative is a NUPPOM as well, but the ARC1 penalty term is applied to parameter vector $\boldsymbol{\beta}_{32}$.

*Simulation sim*1*: case* $D_1 = D_2 = 5$

- $\boldsymbol{\beta}_{10} = (-0.8, 0.4, 1.2, 2.0)'$, $\beta_{11r} = 0.1$, $\beta_{12r} = -0.1$, $r = 1, ..., 4$,
- $\boldsymbol{\beta}_{20} = (-1.5, -0.5, 0.3, 1.4)'$, $\beta_{21c} = 0.1$, $\beta_{22c} = -0.1$, $c = 1, ..., 4$,
- $\boldsymbol{\beta}_{30} = (-0.1, 0.0, 0.0, 0.4, 0.2, 0.2, 0.1, 0.3, 0.3, 0.2, 0.1, 0.1, 0.1, 0.2, 0.1, 0.4)'$,
  $\beta_{31rc} = -0.5$, $\beta_{32rc} = 0$, $r = 1, ..., 4$, $c = 1, ..., 4$.

With this set of parameters (rounded for brevity) the model under the null hypothesis $\beta_{32rc} = 0$, $r = 1, ..., 4$, $c = 1, ..., 4$, is a NUPOM. The model under the alternative is a NUPPOM, with the ARC1 penalty term applied to parameter vector $\boldsymbol{\beta}_{32}$.

*Simulation sim*1*: case* $D_1 = D_2 = 7$

- $\boldsymbol{\beta}_{10} = (-3.3, -2.4, -1.5, -0.7, 1.1, 2.1)'$, $\beta_{11r} = -0.1$, $\beta_{12r} = 0.1$, $r = 1, ..., 6$,
- $\boldsymbol{\beta}_{20} = (-3.5, -2.5, -1.6, -0.8, 0.9, 2.9)'$, $\beta_{21c} = -0.1$, $\beta_{22c} = 0.1$, $c = 1, ..., 6$,
- $\boldsymbol{\beta}_{30} = (3.6, 2.9, 2.7, 2.3, 1.8, 2.1, 3.2, 2.9, 2.6, 2.2, 1.8, 2.0, 2.7, 2.4, 2.1, 1.7, 1.5,$ $1.3, 2.6, 2.1, 1.7, 1.5, 1.3, 1.2, 3.9, 2.2, 1.6, 1.2, 1.2, 1.0, 3.7, 2.3, 1.6, 1.2, 1.1, 1.2)'$, $\beta_{31rc} = 0.1$, $\beta_{32rc} = 0$, $r = 1, ..., 6$, $c = 1, ..., 6$.

With this set of parameters (rounded for brevity) the model under the null hypothesis $\beta_{32rc} = 0$, $r = 1, ..., 6$, $c = 1, ..., 6$, is a NUPOM. The model under the alternative is a NUPOM with the ARC1 penalty term applied to parameter vector $\boldsymbol{\beta}_{32}$. The parameters chosen for the intercepts correspond to the observed global log odds and global odds ratios of the British male occupational study data of Section 6, with an adjustment of 0.0001 for sampling zeros in the original table.

For the three cases outlined in scheme *sim*1, Figure 1 shows some selected histograms of the LR_P simulated distribution with overimposed the theoretical χ^2 with df degrees of freedom that depend on λ. Sample sizes are n=200, 500, and 1000 and correspond to the cases with 3, 5, and 7 levels (*ncat*) per response. For brevity, only the histograms for a reduced set of λ values are reported. The complete set of results for *sim*1 is reported in Supplementary Material. Figure 2 shows the same scheme of Figure 1 but with sample sizes n increased to 500, 1000, and 2000, respectively for *ncat*=3, 5, and 7.

Although the number m of pseudo-samples for this simulation was fixed to 1000, their effective number is reduced because it was based on those models that successfully converged.

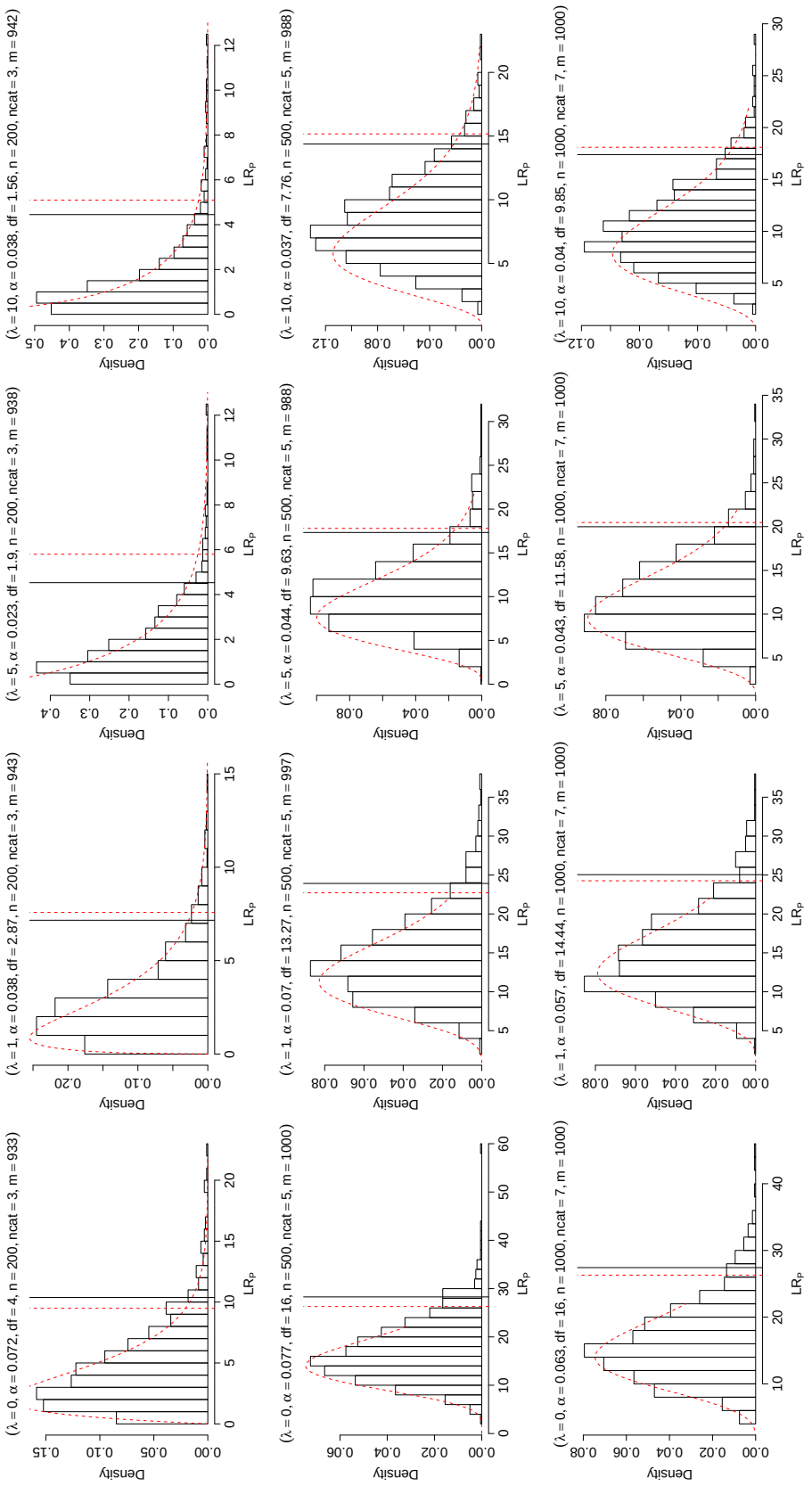

Figure 1: Histograms of the simulated distribution of LR_P from the scheme *sim*1 for varying λ (left to right) and number of response levels *ncat* (up to down). Sample sizes are n=200, 500, and 1000 and vary according to *ncat*, while m indicates the number of valid pseudo-samples. The overimposed dashed curve is a χ^2 with degrees of freedom df that depends on λ. The vertical dashed and continuous lines are in correspondence of the $95th$ percentile of the theoretical and the empirical distribution, respectively.

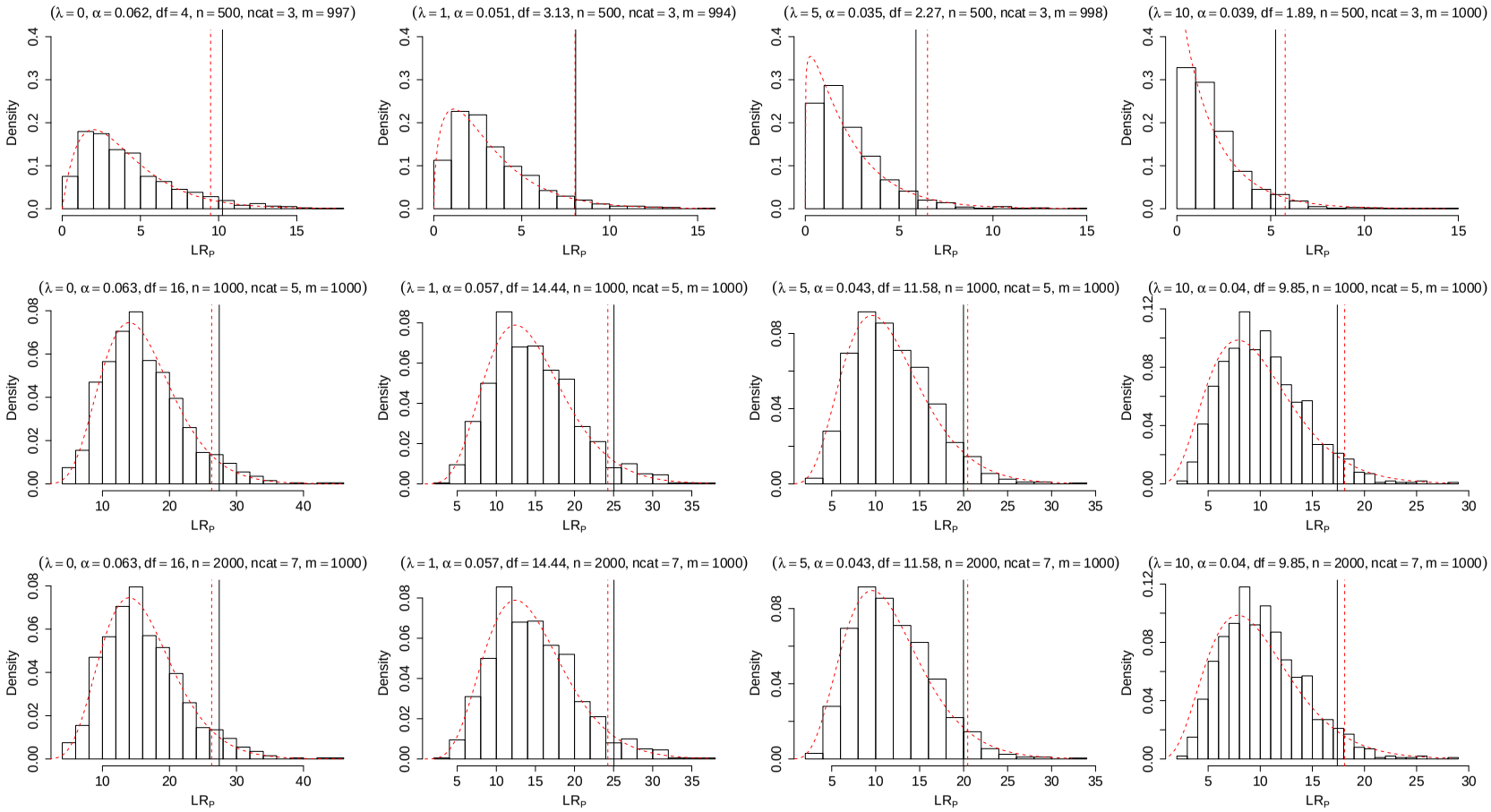

Figure 2: Histograms of the simulated distribution of LR_P from the scheme *sim*1 for varying λ (left to right) and number of response levels *ncat* (up to down). Sample sizes are n=500, 1000 and 2000 and vary according to *ncat*, while m indicates the number of valid pseudo-samples. The overimposed dashed curve is a χ^2 with degrees of freedom (df) that depends on λ. The vertical dashed and continuous lines are in correspondence of the $95th$ percentile of the theoretical and the empirical distribution, respectively.

Overall, the complete set of results for this simulation scheme (reported in the Supplementary Materials) shows a good approximation for larger n and smaller λ.

4.2 Simulation scheme 2

In the second simulation scheme (*sim*2), the sampling distribution of the LR_P statistic is simulated under the null hypothesis in Section 3.2. In all scenarios, the reduced model is a NUPOM and the null hypothesis fixes β_{32rc} to a unique value. The model under the alternative is a NUPPOM as β_{32rc} is unconstrained. In both models, ARC1 is applied to their association intercepts $\boldsymbol{\beta}_{30}$ using the same value of the penalty parameter λ. The parameters chosen to generate the simulation depend on the number of response levels. In particular, these are specified as follows:

*Simulation sim*2*, case* $D_1 = D_2 = 3$

- $\boldsymbol{\beta}_{10} = (-0.5, 0.5)'$, $\boldsymbol{\beta}_{11} = (0.2, 0.2)'$, $\boldsymbol{\beta}_{12} = (-0.3, 0.3)'$,
- $\boldsymbol{\beta}_{20} = (-0.1, 0.6)'$, $\boldsymbol{\beta}_{21} = (0.3, 0.3)'$, $\boldsymbol{\beta}_{22} = (-0.2, 0.4)'$,
- $\boldsymbol{\beta}_{30} = (1.5, 2, 2.5, 3)'$, $\boldsymbol{\beta}_{31} = (-.2, -.2, -.2, -.2)'$, $\boldsymbol{\beta}_{32} = (0.5, 0.5, 0.5, 0.5)'$.

With this set of parameters the null hypothesis is $\beta_{32rc} = 0.5$, $r = 1, 2$, $c = 1, 2$.

*Simulation sim*2*, case* $D_1 = D_2 = 5$

- $\boldsymbol{\beta}_{10} = (-0.8, 0.4, 1.2, 2.0)'$, $\beta_{11r} = 0.1$, $\beta_{12r} = -0.1$, $r = 1, ..., 6$,
- $\boldsymbol{\beta}_{20} = (-1.5, -0.5, 0.3, 1.4)'$, $\beta_{21c} = 0.1$, $\beta_{22c} = -0.1$, $c = 1, ..., 4$,
- $\boldsymbol{\beta}_{30} = (-0.1, 0.0, 0.0, 0.4, 0.2, 0.2, 0.1, 0.3, 0.3, 0.2, 0.1, 0.1, 0.1, 0.2, 0.1, 0.4)'$, $\beta_{31rc} = 0.2$, $\beta_{32rc} = 0.4$, $r = 1, ..., 4$, $c = 1, ..., 4$.

With this set of parameters (rounded for brevity to the first decimal point) the null hypothesis is $\beta_{32rc} = 0.4$, $r = 1,...,4$, $c = 1,...,4$,.

Simulation sim2, case $D_1 = D_2 = 7$

- $\boldsymbol{\beta}_{10} = (-3.1 - 2.3 - 1.4 - 0.71.12.1)'$, $\beta_{11r} = -0.1$, $\beta_{12r} = 0.1$, $r = 1,...,6$,
- $\boldsymbol{\beta}_{20} = (-3.3, -2.4, -1.5, -0.8, 0.9, 2.0)'$, $\beta_{21c} = -0.1$, $\beta_{22c} = 0.1$, $c = 1,...,6$,
- $\boldsymbol{\beta}_{30} = (3.2, 2.6, 2.5, 2.0, 1.4, 1.4, 2.8, 2.6, 2.4, 2.0, 1.4, 1.4, 2.4, 2.2, 1.9, 1.6, 1.3,$ $1.1, 2.2, 1.9, 1.6, 1.4, 1.2, 1.1, 1.8, 1.5, 1.3, 1.1, 1.1, 0.9, 1.5, 1.4, 1.2, 1.0, 1.0, 1.1)'$, $\beta_{31rc} = -0.1$, $\beta_{32} = 0.4$, $r = 1,...,6$, $c = 1,...,6$.

With this set of parameters (rounded) the null hypothesis is $\beta_{32rc} = 0.4$, $r = 1,...,6$, $c = 1,...,6$,. The parameters chosen for the intercepts correspond to the observed global log odds and global odds ratios of the British male occupational study data of Section 6, with an adjustment of 0.001 for sampling zeros in the original table.

For the simulation scheme 2, Figure 3 shows some selected histograms of the LR_P simulated distribution with overimposed the theoretical χ^2 with 3, 15, and 35 degrees of freedom, and for *ncat*= 3, 5, and 7 levels, respectively. Sample sizes are *n*=200, 500, and 1000 and correspond to *ncat*=3, 5, and 7, as well. The complete set of results for *sim*2 is reported in Supplementary Material.

Figure 4 shows the same scheme of Figure 3 but with sample sizes n increased to 500, 1000, and 2000, respectively for *ncat*=3, 5, and 7.

Overall, the histograms for *sim*2 shows a good approximation of the theoretical χ^2 to the LR_P sampling distribution for larger n and smaller λ. On the contrary, the approximation

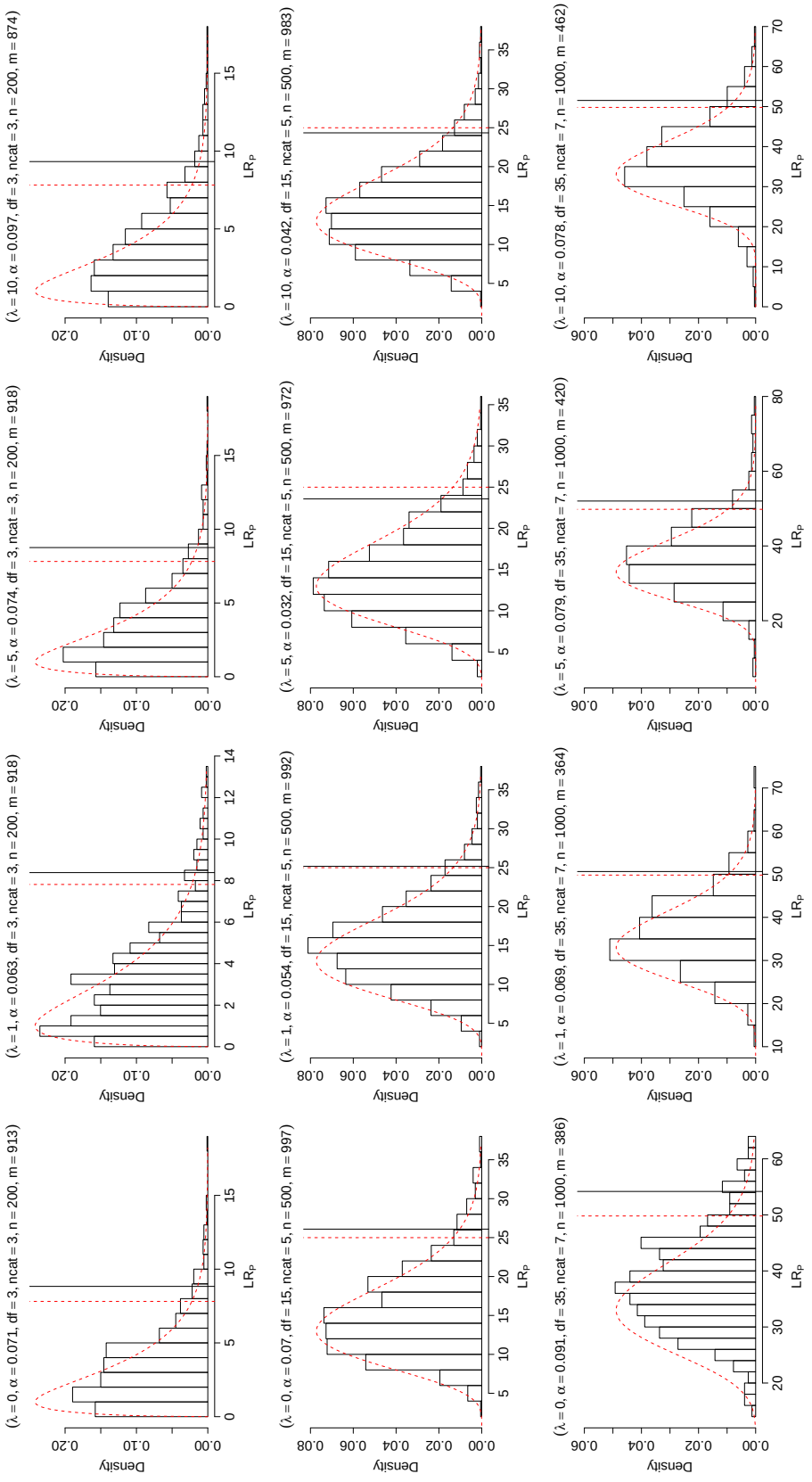

Figure 3: Histograms of the simulated distribution of LR_P from the scheme *sim2* for varying λ (left to right) and number of response levels (up to down). Sample sizes are n=200, 500, and 1000 and vary according to *ncat*, while m indicates the number of valid pseudo-samples. The overimposed dashed curve is a χ^2 distribution. The vertical dashed and continuous lines are in correspondence of the 95th percentile of the theoretical and the empirical distribution, respectively.

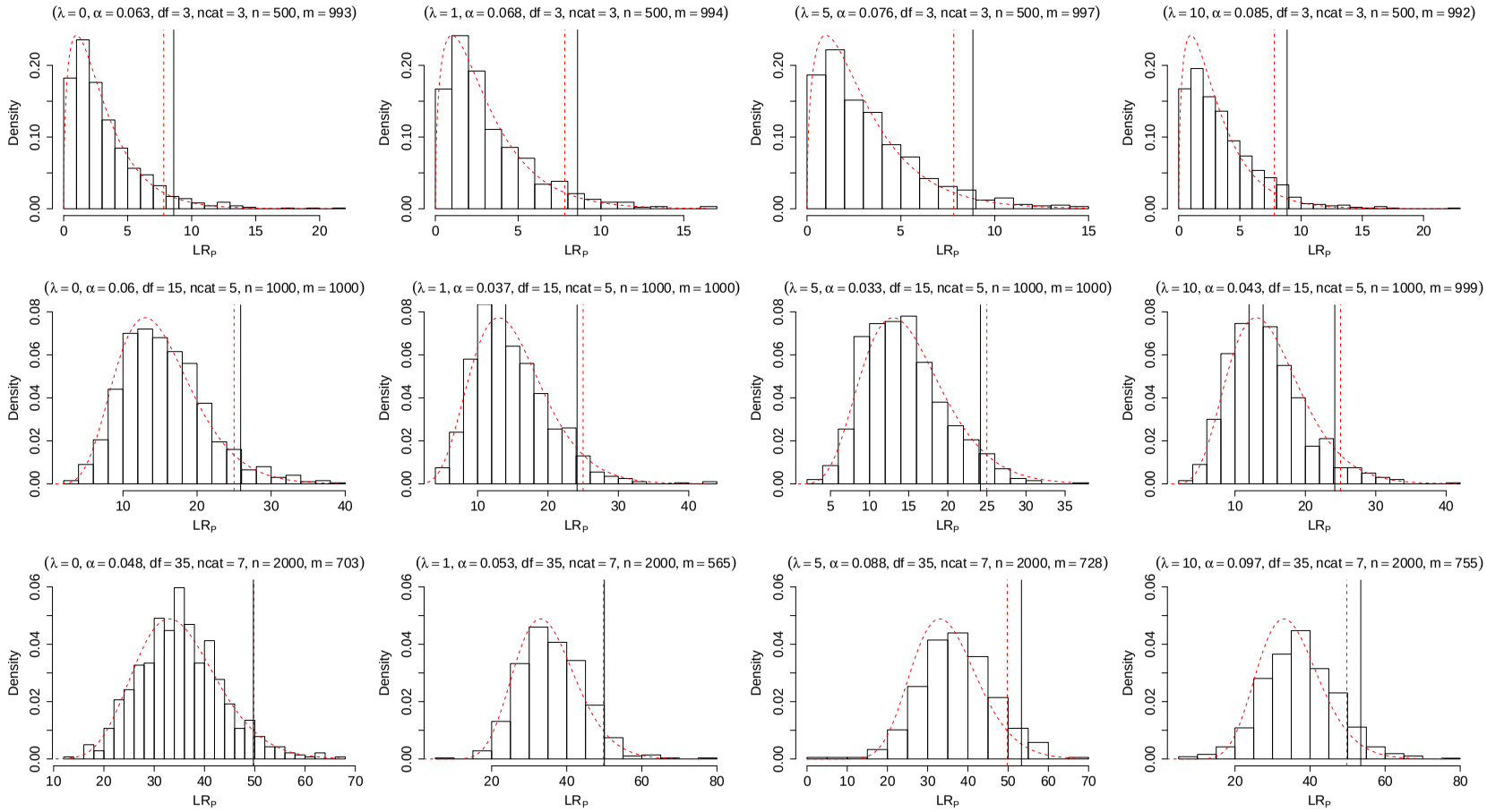

Figure 4: Histograms of the simulated distribution of LR_P from the scheme *sim*2 for varying λ (left to right) and number of response levels (up to down). Sample sizes are n=500, 1000, and 2000 and vary according to *ncat*, while m indicates the number of valid pseudo-samples. The overimposed dashed curve is a χ^2 distribution. The vertical dashed and continuous lines are in correspondence of the 95th percentile of the theoretical and the empirical distribution, respectively.

is very poor especially for $\lambda > 10$ (see also Supplementary Material).

5 Evaluating the performance of penalized estimates

In order to evaluate the potential of smoothed estimates, a simulation study for the case with three response levels and one continuous covariates is carried out. Three sample sizes are considered n=200, 500 and 1000, whereas the number of samples is taken to be $m = 1000$. The simulations are generated from a *NUNPOM* with the following parameters:

- $\boldsymbol{\beta}_{10} = (-0.6, 0.6)'$, $\boldsymbol{\beta}_{11} = (0.3, -0.3)'$,
- $\boldsymbol{\beta}_{20} = (-0.6, 0.6)'$, $\boldsymbol{\beta}_{21} = (-0.3, 0.3)'$,
- $\boldsymbol{\beta}_{30} = (2.6, 2.4, 2.0, 1.7)'$, $\boldsymbol{\beta}_{31} = (-0.4, 0.2, -0.5, 0.5)'$.

The values $x_i, i = 1, ..., n$ of the covariate were drawn from a $U(-1, 1)$. Thus, given the model formula (2.6) and the inversion method (2.4) we found the $n \times (D_1 D_2)$ probability matrix Π, in which each row $\boldsymbol{\pi}_i'$ represents the probability vector for the ith observation. Then, the responses were drawn from a multinomial distribution with probability vector $\boldsymbol{\pi}_i'$. The *UPOM* was compared to the *NUNPOM*, for which ARC1 has been used in combination with (2.16). Penalization parameters for ARC1, that is λ_1, λ_2 and λ_3 were chosen equal to a single value λ, varying in the set $\{0, 1, 10, 100, ...\}$. The results of the simulation were evaluated by the number of Fisher scoring successes, by the AIC (defined in Appendix A), by the loss functions:

Mean squared error loss:

$$MSEL = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{D_1 D_2}(\pi_{ir} - \hat{\pi}_{ir})^2, \tag{5.1}$$

Mean relative squared error loss:

$$MRSEL = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{D_1 D_2}\frac{(\pi_{ir} - \hat{\pi}_{ir})^2}{\pi_{ir}}, \tag{5.2}$$

Mean entropy or Kullback-Leibler loss:

$$MEL = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{D_1 D_2}\pi_{ir}\log\left(\frac{\pi_{ir}}{\hat{\pi}_{ir}}\right), \tag{5.3}$$

and by the overall Relative Bias:

$$RBIAS = \frac{1}{FSS}\sum_{f=1}^{FSS}\sum_{k=1}^{3}\sum_{q=1}^{Q_k}\frac{|\hat{\beta}_{fkq} - \beta_{fkq}|}{\beta_{fkq}}, \tag{5.4}$$

where FSS is the number of Fisher scoring successes. $\hat{\beta}_{fkq}$ and β_{fkq} represent the qth estimated parameter and the qth true parameter, respectively, in the kth model equation of the fth simulation. In our example, $Q_1 = Q_2 = 4$ while $Q_3 = 8$.

Before setting λ to some value greater than zero, some attempts to estimate the model were made by reducing the step length l_{step} of the iterative algorithm (see Appendix A), and in some case the *NUNPOM* was fitted. Actually, this step length reduction is automatically performed with the 'pblm' R package provided along with this paper. The *UPOM* was fitted in all simulations. The results are reported in Table 1, while Table 2 reports the mean estimated βs.

Table 1: Comparison $UPOM$ vs a $NUNPOM$ with ARC1, in terms of 1000 simulations generated under the NUNPOM assumption, with $ncat = 3$ for both responses, sample sizes of n=200, 500, and 1000, and $\lambda = 0, 1, 10, 100$. Mean values of loss functions, *AIC*, and relative bias are reported as well as the number of Fisher scoring successes.

n	*Model*	λ	*MSEL*	*MRSEL*	*MEL*	*AIC*	*RBIAS*	*FSS*
200	NUNPOM	0	0.0081	0.0786	0.0425	808.02	2.50	298
200	NUNPOM	1	0.0076	0.0735	0.0373	806.16	3.02	514
200	NUNPOM	10	0.0080	0.0843	0.0397	803.66	5.05	877
200	NUNPOM	100	0.0114	0.1521	0.0596	810.18	7.23	992
200	UPOM	-	0.0130	0.1871	0.0686	809.73	7.86	1000
500	NUNPOM	0	0.0032	0.0320	0.0161	1991.39	1.20	445
500	NUNPOM	1	0.0031	0.0315	0.0156	1989.65	1.93	608
500	NUNPOM	10	0.0033	0.0360	0.0168	1987.57	4.04	904
500	NUNPOM	100	0.0061	0.0850	0.0325	2001.85	6.61	1000
500	UPOM	-	0.0092	0.1497	0.0499	2019.30	7.94	1000
1000	NUNPOM	0	0.0017	0.0160	0.0080	3957.54	0.5	633
1000	NUNPOM	1	0.0016	0.0161	0.0080	3954.82	1.15	743
1000	NUNPOM	10	0.0018	0.0193	0.0091	3953.41	3.36	931
1000	NUNPOM	100	0.0037	0.0506	0.0200	3973.60	5.80	1000
1000	UPOM	-	0.0085	0.1492	0.0474	4028.52	7.89	1000

In the simulation with n=200 and $\lambda = 0$, Fisher scoring failed in 702 out of 1000 simulations, while almost all models were successfully estimated when $\lambda = 100$. Observe that all the mean loss functions and *AIC* for the $NUNPOM$ are smaller than the $UPOM$ ones, as it should be. The $NUNPOM$ without penalization (i.e. $\lambda = 0$) has the smallest loss functions values, but also the greatest *AIC* value among the $NUNPOMs$.

Table 2: Comparison *UPOM* vs *NUNPOM* with ARC1 in terms of mean parameters of 1000 simulations, of sizes n=200, 500 and 1000, generated under the NUNPOM assumption with $ncat$=3 for both responses.

n	Model	λ	β_{101}	β_{102}	β_{111}	β_{112}	β_{201}	β_{202}	β_{211}	β_{212}	β_{301}	β_{302}	β_{303}	β_{304}	β_{305}	β_{312}	β_{313}	β_{314}
200	TRUE MODEL	0	-0.60	0.60	0.30	-0.30	-0.60	0.60	-0.30	0.30	2.60	2.40	2.00	1.70	-0.40	0.20	-0.50	0.50
200	NUNPOM	0	-0.61	0.60	0.28	-0.27	-0.62	0.60	-0.27	0.29	2.68	2.61	2.17	1.73	-0.40	-0.08	-0.27	0.48
200	NUNPOM	1	-0.60	0.61	0.28	-0.25	-0.61	0.61	-0.28	0.27	2.69	2.51	2.12	1.69	-0.25	-0.03	-0.15	0.39
200	NUNPOM	10	-0.59	0.59	0.21	-0.17	-0.59	0.59	-0.26	0.14	2.72	2.46	2.14	1.69	0.03	0.07	0.07	0.22
200	NUNPOM	100	-0.59	0.59	0.10	0.01	-0.60	0.60	-0.16	-0.07	2.69	2.50	2.12	1.62	0.14	0.14	0.14	0.16
200	UPOM	-	-0.59	0.59	0.06	-0.61	0.61	-0.13	2.65	2.50	2.11	1.59	0.15	-0.59	0.59	0.06	-0.61	0.61
500	NUNPOM	0	-0.61	0.61	0.29	-0.29	-0.60	0.61	-0.29	0.28	2.63	2.48	2.05	1.70	-0.39	0.07	-0.39	0.52
500	NUNPOM	1	-0.61	0.60	0.29	-0.28	-0.60	0.60	-0.31	0.28	2.64	2.44	2.04	1.69	-0.34	0.01	-0.27	0.46
500	NUNPOM	10	-0.60	0.60	0.26	-0.24	-0.59	0.59	-0.31	0.21	2.70	2.38	2.07	1.65	-0.10	-0.01	-0.02	0.27
500	NUNPOM	100	-0.60	0.60	0.14	-0.05	-0.58	0.58	-0.21	-0.02	2.69	2.43	2.08	1.60	0.09	0.10	0.10	0.14
500	UPOM	-	-0.61	0.61	0.07	-0.59	0.58	-0.15	2.63	2.45	2.07	1.56	0.11	-0.61	0.61	0.07	-0.59	0.58
1000	NUNPOM	0	-0.60	0.60	0.29	-0.30	-0.60	0.61	-0.30	0.29	2.60	2.43	2.01	1.69	-0.40	0.13	-0.48	0.50
1000	NUNPOM	1	-0.60	0.60	0.30	-0.29	-0.60	0.60	-0.31	0.28	2.62	2.41	2.01	1.69	-0.36	0.08	-0.39	0.46
1000	NUNPOM	10	-0.60	0.59	0.28	-0.27	-0.60	0.59	-0.32	0.24	2.68	2.37	2.03	1.66	-0.19	-0.04	-0.13	0.31
1000	NUNPOM	100	-0.59	0.59	0.18	-0.12	-0.58	0.58	-0.25	0.06	2.70	2.40	2.05	1.63	0.05	0.07	0.06	0.15
1000	UPOM	-	-0.60	0.60	0.07	-0.59	0.59	-0.15	2.60	2.44	2.04	1.56	0.10	-0.60	0.60	0.07	-0.59	0.59

As expected (Table 2), especially for the NUNPOM parameters involved in the penalization, the larger λ values the larger the parameter bias. However, overall, the bias appears to be smaller than the UPOM one. Further, the bias decreases as n increases.

6 Applications to real data sets

6.1 The British male occupational status data set

Consider the data on occupational status (OS) of a sample of British males from Goodman (1979), where fathers and their sons were cross-classified according to the occupational status using seven ordered categories. The data are reported in Table 3.

Table 3: Cross-classification of British males according to the occupational status.

Father's status	Subject's status						
	1	2	3	4	5	6	7
1	50	19	26	8	18	6	2
2	16	40	34	18	31	8	3
3	12	35	65	66	123	23	21
4	11	20	58	110	223	64	32
5	14	36	114	185	714	258	189
6	0	6	19	40	179	143	71
7	0	3	14	32	141	91	106

Several authors have re-analyzed such data. For example, Lapp et al. (1998) compare the Goodman RC and Dale models in terms of goodness-of-fit. We further re-analyze the data

by fitting the BOLM with ARC2. The aim of the application is to show the advantages of our proposal when compared to the existing alternatives.

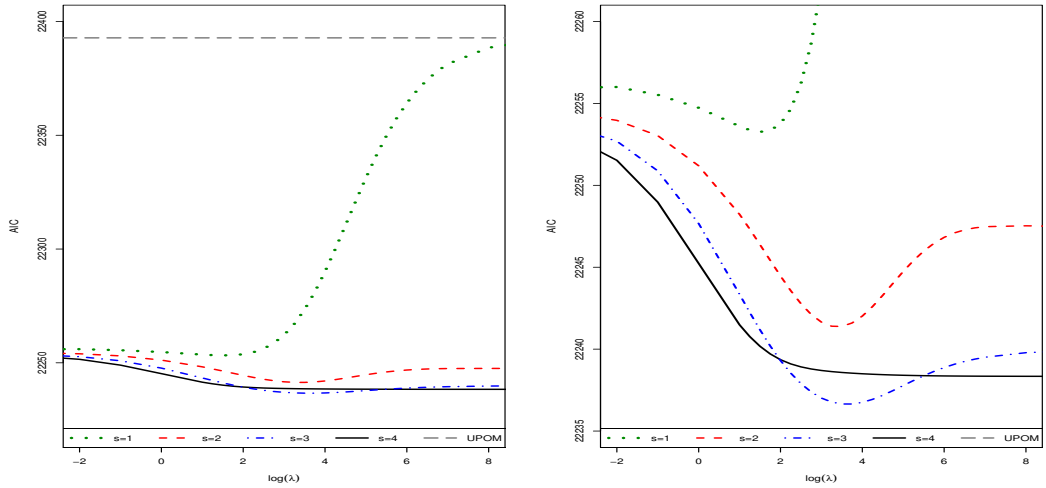

Figure 5: OS data set: the left plot shows the AIC for varying smoothing parameter λ (in log scale) for a NUPOM using the penalty term ARC2 on the association intercepts. Different orders s of penalization are used assuming $s_3 = s_4 = s$. For $\log\lambda \to +\infty$, $s = 1$ (dotted line) the model AIC will tend to the UPOM level; $s = 2$ (dashed line), $s = 3$ (dotted-dashed line) and $s = 4$ (continuous line) correspond to AIC of models tending to a polynomial from first to third degree, respectively. The plot on the right shows the detail of the most critical interval, where the AIC is minimized for $\log\lambda = 4$ and $s = 3$.

The saturated model for the joint distribution involves 48 parameters: 6 global logits for each marginal and 36 log-GORs. Since the interest is in modelling the association structure, the focus is on the 36 log-GORs only. Figure 5 shows the AIC for the NUPOM for varying smoothing parameter and different orders of penalization.

Due to the symmetry of the association structure, the penalization orders of the difference operator s_3 and s_4 are assumed to be equal and indicated by s. The AIC for the model with $s = 1$ tends to the UPOM AIC level for high values of $\log\lambda$. This model is clearly

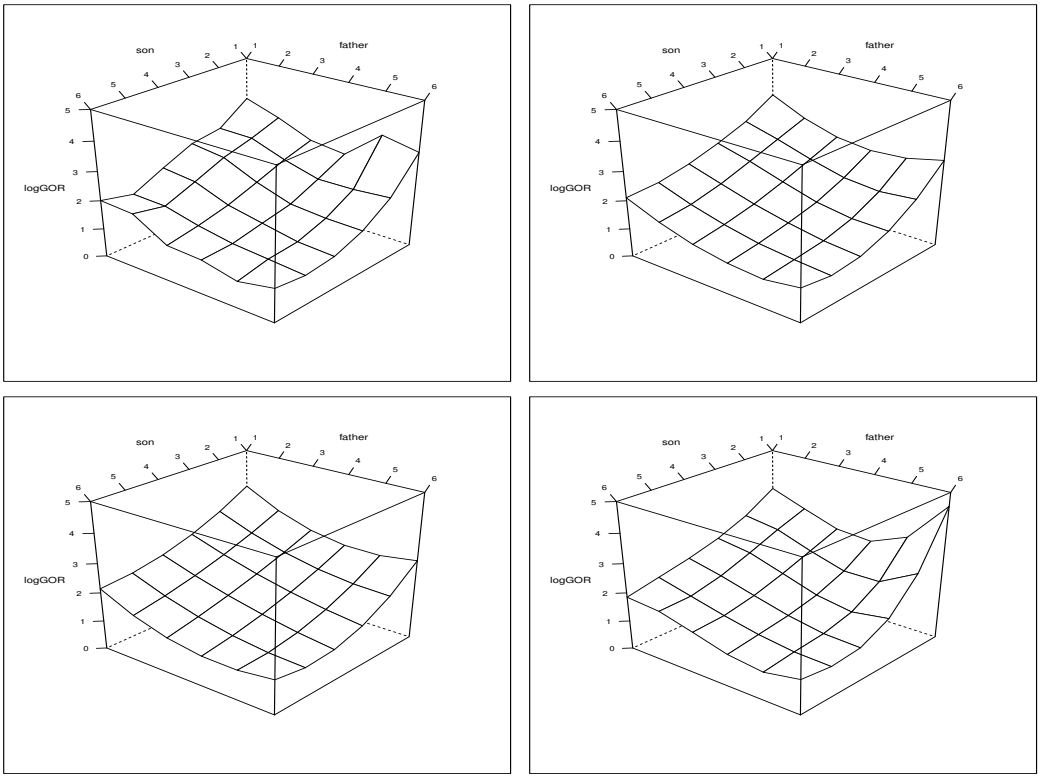

Figure 6: Association structures for the OS data set. In lexicographical order: observed log-GORs (top-left); log-GORs fitted by the interaction of polynomials of second degree with non-integer scores (top-right); log-GORs fitted by the interaction of polynomials of second (bottom-left) and third (bottom-right) degree with integer scores.

inadequate as the models specified by higher s provide smaller AIC whatever $\log\lambda$. The minimum value of AIC is 22236.65, corresponding to $\log(\lambda) = 4$ and $s = 3$. This represents a model with an association structure which tends to a smooth surface defined by row and column interactions of second degree polynomials. On the grounds of *AIC* only, one could choose this model. However, for high values of λ, the models with $s = 3$ $(AIC = 22239.96)$ and $s = 4$ $(AIC = 22238.34)$ provide good fits as well, with a slight evidence in favor of the latter model, corresponding to a polynomial surface of third degree. The AIC differences

of such models, with respect to the minimum AIC, are respectively 3.31 and 1.69, which are quite small (Burnham and Anderson, 2000, p. 48). When it is possible, as in this case, it is preferable to choose a model providing integer and equally-spaced scores, on the grounds of greater interpretability and for the possibility to use classical test statistics whose asymptotic null distributions are well-known. Therefore, we report in Table 4 the results (in terms of *AIC* and deviance G^2) for the four polynomial models evaluated at the largest value of $\log(\lambda) = 15$, along with the independence and saturated models.

Table 4: Model selection based on AIC and deviance G^2 for the OS data set based on a BOLM with penalty term ARC2. The asterisks indicate a non significant difference with the saturated model.

Model	Description	df	AIC	G^2
1	Independence	36	23081.12	897.52
2	Uniform association	35	22392.83	207.22
3	First degree polynomials	32	22247.46	55.85
4	Second degree polynomials	27	22239.96	*38.36
5	Third degree polynomials	20	22238.34	**22.74
6	Saturated	0	22255.60	0.00

*p=0.07, **p=0.3.

Model 1 has been fitted by using the ridge-type penalty term, such that the estimated global log-odds ratios tend to zero for high values of the smoothing parameter. Model 4 provides the most parsimonious but yet acceptable fit, with only 9 estimated parameters and p-value = 0.07. Model 5 estimates only 16 parameters, providing a comparable fit ($G^2 = 22.74$), with a not significant difference with the saturated model (p-value = 0.3). This means the ordinal association structure of occupational status can be well fitted by a polynomial

of second or third degree. Notice that Models 4 and 5 are more parsimonious than the best model found in Lapp *et al.*'s analysis, i.e. the Dale model, including row effects, column effects, and interactions, while maintaining a comparable fit in terms of G^2. The observed structure of global log-odds ratios and the selected polynomial models are graphically showed in Figure 6.

As we can see, the ordinal association structure is always positive, but it decreases as both the social statuses increase. Observe that the second degree polynomials using integer and non-integer scores show very slight differences. Finally, notice that also Lapp *et al.* hypothesized the possibility to fit a "symmetric second degree polynomial" model.

6.2 The liver disease patients data set

The data set consists of 256 patients with a liver disease progression. The two outcomes, both measured on the same day, are the liver biopsy (named *STAGE*), considered the natural gold standard, and a categorized version of transient elastography (*STIFF*) according to cutoffs suggested by Castera et al. (2005), to measure liver stiffness. Both responses have three ordered categories, *STIFF* with levels 1,2 and 3, which correspond to stiffness classes $[0, 7.1)$, $[7.1, 12.5)$ and $[12.5, \infty)$, respectively; as for *STAGE*, the initial five categories F0-F4 have been collapsed as follows: 1, corresponding to the biopsy stage $< F2$, 2 $(= \{F2, F3\})$ and 3 $(= F4)$. Aim of the study is to evaluate the concordance between the outcomes in order to find profiles of "discordant" patients. This is done using the bivariate logistic model, by employing the log global odds as marginal parameters and the log global odds ratio as association measure. For these data, a first analysis with dichotomized responses was made by Calvaruso et al. (2010). Table 5 shows the cross-classification of the responses, ignoring covariates.

Table 5: Marginal Cross-classification of the responses and empirical global log-odds ratios for the liver disease patient data.

	STIFF		
STAGE	1	2	3
1	71	20	0
	(1.72)	($+\infty$)	
2	56	20	8
	(3.18)	(3.31)	
3	8	27	46

From a first look at Table 5, it is possible to notice a positive association between the responses, even if there are many discordant patients, mainly the fifty-six in $STAGE = 2, STIFF = 1$. Among the covariates, the patient's gender (*SEX*), age (*AGE*), alanine aminotransferase (*ALT*) measured in U/L and platelet (*PLT*) levels measured in 10^3 mmc, are considered. For modelling purposes, the covariates were centered with respect to their means and, after a backward selection, we considered the following sets of variables:

$$\mathscr{P}_1^0 = \{\text{Intercept}_{\overline{123}}, SEX_{123}, AGE_{123}, ALT_{123}, PLT_{1\bar{2}3}\},$$

$$\mathscr{P}_2^0 = \{\text{Intercept}_{\overline{123}}, SEX_{123}, AGE_{123}, ALT_{123}, PLT_{123}\},$$

$$\mathscr{P}_3^0 = \{\text{Intercept}_{\overline{123}}, SEX_{123}, AGE_{123}, ALT_{123}, PLT_{1\bar{2}3}\},$$

$$\mathscr{P}_4^0 = \{\text{Intercept}_{\overline{123}}, SEX_{123}, AGE_{123}, ALT_{123}, PLT_{123}\},$$

$$\mathscr{P}_5^0 = \{\text{Intercept}_{\overline{123}}, AGE_{123}, ALT_{123}, PLT_{1\bar{2}3}\},$$

$$\mathscr{P}_6^0 = \{\text{Intercept}_{\overline{123}}, AGE_{123}, ALT_{123}, PLT_{1\bar{2}}\},$$

$$\mathscr{P}_7^0 = \{\text{Intercept}_{\overline{123}}, AGE_{12}, ALT_{123}, PLT_{1\bar{2}}\},$$

where $\text{Intercept}_{\overline{123}}$ indicates that the marginal and association intercepts are category dependent, whereas $\text{Intercept}_{\overline{123}}$ indicates that the intercepts for the association are category independent. To indicate that variable *AGE* is included both in marginal and association predictors, we use AGE_{123}, whereas AGE_{12} indicates that such variable is included only in the marginal predictors. Computational problems had arisen when we tried to estimate non-uniform association models. Such problems were overcome by regularizing the parameter space of the association intercepts. In particular, the ARC1 penalty term was employed, with smoothing parameter $\lambda_3 = 0.5$, which is the value that minimizes the AIC. This value was found through the two-step procedure described in Section 2.1.1. By considering the results from the simulation in Section 3, we decided to use a χ^2 distribution to approximate the LR_P asymptotic distribution. The results of model selection are reported in Table 6.

Table 6: Model selection based on the *AIC* and the LR_P statistic for the liver disease patients data using the ARC1 penalty term.

Model	Description	n. par.	AIC	vs	LR_P	df	p-value
1	$NUPPOM(\mathscr{P}_1^0)$	20.2	877.41	-	-	-	-
2	$NUPOM(\mathscr{P}_2^0)$	19.2	879.86	1	4.58	1	0.032
3	$UPPOM(\mathscr{P}_3^0)$	18	895.72	1*	19.08	2	< 0.001
4	$UPOM(\mathscr{P}_4^0)$	17	898.21	1*	23.57	3	< 0.001
5	$NUPPOM(\mathscr{P}_5^0)$	18.2	872.50	1	1.33	3	0.745
6	$NUPPOM(\mathscr{P}_6^0)$	17.2	870.69	5	0.28	1	0.595
7	$NUPPOM(\mathscr{P}_7^0)$	16.2	870.72	6	1.53	1	0.217
7	-	-	-	1	3.04	5	0.694

*Obtained by rounding the degrees of freedom of Model 1 to 20.

For each model that we have selected, the table reports its description, the number of estimated parameters and the AIC. The next columns refer to the comparisons between nested models, specified by the column headed "vs". The last three columns report the results of such comparison in terms of penalized log-likelihood ratio statistic, along with degrees of freedom and p-values. Before proceeding to variable selection, the hypothesis $UPOM$, versus alternatives $UPPOM$, $NUPOM$, and $NUPPOM$ were checked for all variables. The table reports the comparisons for Models 1-4. Model 1 is the most complex model we have considered, a $NUPPOM$ defined on set $\mathscr{P}_1^0$. This model assumes that the effect of variable PLT on $STIFF$ depends on the categories of $STIFF$. Models 2-4 represent hypotheses of uniform association and/or (partially) proportional odds, and these models are compared to Model 1, for which none of these hypotheses holds. Although the LR_P test for model comparison is approximated, some results seem to be clear. For example, because the difference between model 1 and 3 (or 4) is highly significant, the hypothesis of $UPPOM$ (or $UPOM$) does not hold. Models 5-7 concern a backward model selection starting from Model 1. The last row reports the comparison between Models 7 and 1, for which the difference between the starting model and the final model is not significant (p-value=0.711). In model 7, variable ALT is the only one which has significant (global) effect for the association model. By AIC, the model with the best trade-off between goodness-of-fit and parsimony is still Model 7. Figure 7 shows the comparison between the simulated distribution of LR_P for the comparison between the models in Table 6.

Models 3 and 4 are not involved in this simulation for the reason explained above. α is the real significance level obtained using the 95th percentile of the χ^2 distribution (dashed curve) with respect to the empirical one, while m* indicates the number of valid pseudo-samples from the initial m=1500. Vertical dashed and continuous lines are in correspondence of the 95th percentile of the theoretical and the empirical distribution, respectively.

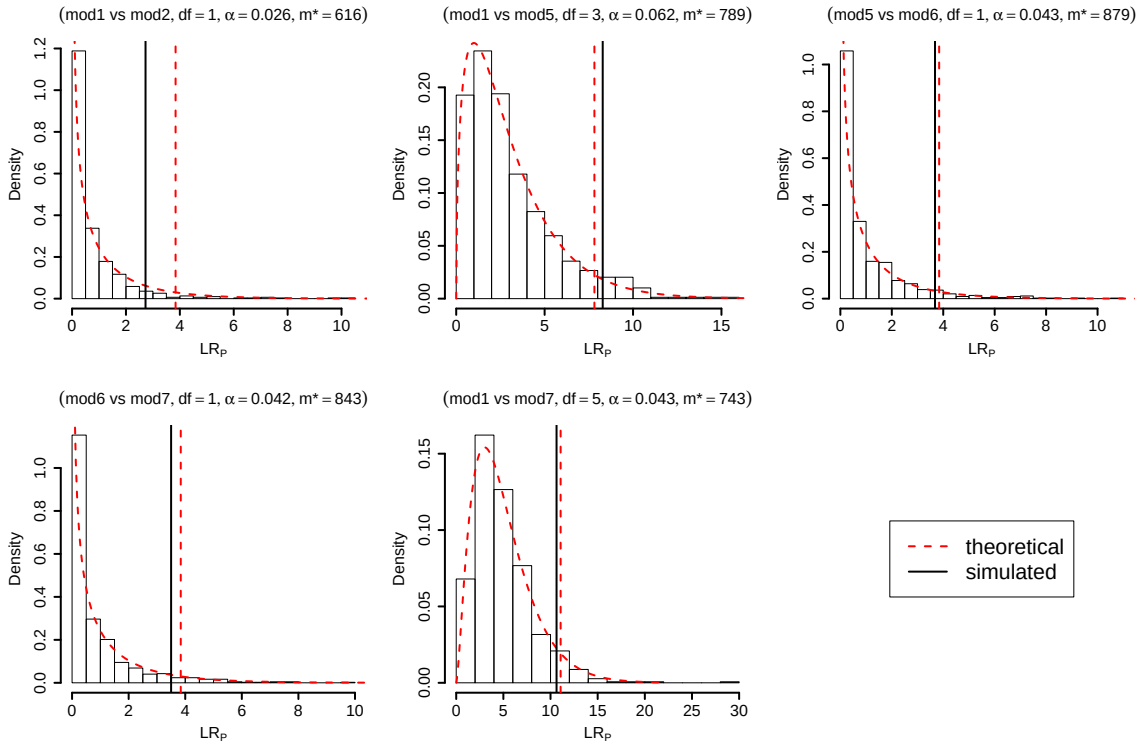

Figure 7: Simulated distribution of LR_P for the comparison between the models in Table 6. The overimposed dashed curve is a χ^2 distribution. The vertical lines are in correspondence of the 95th percentile of the theoretical and the empirical distribution, respectively.

Overall, there is a good correspondence when the χ^2 distribution is used to approximate the simulated distribution, except for the comparison between Model 1 and Model 2. For this comparison, the real significance level, when using the 95th percentile of the χ^2_1, is $\alpha = 0.03$, resulting to be more conservative. Estimates for the final Model 7 are reported in Table 7. Variables *ALT*, *AGE* and *PLT* are significant in both marginal outcomes. In particular, the platelet level has a category-dependent effect for *STIFF* which is higher for the log global odds 1-2 than 3. In particular, a patient at older age, higher *ALT* and lower *PLT* values is more at risk of having a greater liver stiffness than a patient with mean values. In addition, the *ALT* effect for *STIFF* is about twice as strong as for biopsy stage. The

effect of ALT in the association is significant, and considering the intercepts values as well, higher ALT values imply a global reduction of the association, especially for individuals in class $STIFF < 7.1$ and $STAGE = 1$.

Table 7: Estimates for Model 7.

response	variable	estimate	se	z	p.value
STAGE	Intercept 1	-0.7404	0.1415	-5.233	< 0.001
	Intercept 2	0.8786	0.1451	6.053	< 0.001
	ALT	-0.0047	0.0018	-2.649	0.008
	AGE	-0.0404	0.0104	-3.878	< 0.001
	PLT	0.0093	0.0021	4.444	< 0.001
STIFF	Intercept 1	0.1003	0.1377	0.728	0.466
	Intercept 2	1.7868	0.1999	8.938	< 0.001
	ALT	-0.0090	0.0019	-4.726	< 0.001
	AGE	-0.0421	0.0110	-3.828	< 0.001
	PLT 1	0.0087	0.0024	3.675	< 0.001
	PLT 2	0.0154	0.0029	5.243	< 0.001
ASSOCIATION	Intercept 1	1.4934	0.3336	4.473	< 0.001
	Intercept 2	3.9066	0.7828	4.998	< 0.001
	Intercept 3	2.8217	0.3978	7.118	< 0.001
	Intercept 4	2.8770	0.4336	6.635	< 0.001
	ALT	-0.0081	0.0036	-2.244	0.025

7 Discussion

We have shown how to fit a BOLM by penalized ML estimation with some penalty terms for a "vertical penalization", that is across response levels. Particular emphasis on the terms ARC1 and ARC2 has been given. The motivation for our approach is, on one hand, its flexibility in modelling situations in which ML estimation by Fisher scoring appears somewhat difficult and, on the other hand, the possibility to consider the fit of a $NUPPOM$, which lies between a $UPOM$, which may give a poor fit, and a $NUNPOM$, often less useful and somewhat more complicated to estimate than a $UPOM$. The penalized log-likelihood ratio LR_P statistic has been considered to check the hypothesis that certain effects are category independent. To our knowledge, the asymptotic distribution of LR_P for the considered hypothesis is not known, though we have shown by simulation that for relatively small smoothing values the χ^2 may be a good approximation. However, as far as the distributional properties of penalized likelihood ratio test-statistics are concerned, further investigations are necessary. The potential of penalized estimates by penalty term ARC1 has been shown by simulation and by an application to an original data set. In addition, the BOLM has been fitted using the penalty term ARC2 to a literature data set for comparison with the alternative Dale and Goodman RC models, showing parsimony while preserving a satisfactory the goodness-of-fit. In some sense, ARC2 generalizes ARC1, permitting to fit restricted versions of the Dale model, by inserting row or column effects, but also polynomial effects models, with scores chosen by data.

All codes and applications have been included into Supplementary Material along with the 'pblm' R package, used in this paper for all computations. Further, the package (not on CRAN at time of writing) permits to fit additive BOLMs using P-splines.

Appendix A: penalized maximum likelihood estimation

Let $\partial l/\partial \boldsymbol{\pi}_i = \text{diag}(\boldsymbol{\pi}_i)^{-1}\boldsymbol{y}_i$, $\partial \boldsymbol{\pi}_i/\partial \boldsymbol{\eta}_i = (\mathbf{C}'\mathbf{D}_i^{-1}\mathbf{L})^{-1}$ and $\partial \boldsymbol{\eta}_i/\partial \boldsymbol{\beta} = \mathbf{X}_i$. By using the chain rule, the first derivative of the penalized log likelihood with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial l_P}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} \frac{\partial l}{\partial \boldsymbol{\pi}_i}\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} - \mathbf{P}\boldsymbol{\beta},$$

the *penalized score function* is

$$\boldsymbol{s}_P(\boldsymbol{\beta};\boldsymbol{y}_i) = \sum_{i=1}^{m} [(\mathbf{C}'\mathbf{D}_i^{-1}\mathbf{L})^{-1}\mathbf{X}_i]'\text{diag}(\boldsymbol{\pi}_i)^{-1}\boldsymbol{y}_i - \mathbf{P}\boldsymbol{\beta},$$

and the *penalized Fisher matrix* is

$$\mathbf{F}_P(\boldsymbol{\beta}) = \sum_{i=1}^{m} n_i\mathbf{X}_i'(\mathbf{L}'\mathbf{D}_i'^{-1}\mathbf{C})^{-1}\text{diag}(\boldsymbol{\pi}_i)^{-1}(\mathbf{C}'\mathbf{D}_i^{-1}\mathbf{L})^{-1}\mathbf{X}_i + \mathbf{P}.$$

Using these formulas, the $(k+1)$th iteration of the Fisher scoring is $\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + l_{step}\mathbf{F}_P(\hat{\boldsymbol{\beta}}^{(k)})^{-1}\boldsymbol{s}_P(\hat{\boldsymbol{\beta}}^{(k)})$, where l_{step} is a positive scalar representing the step length. Since the iterative procedure may produce incompatible $\boldsymbol{\beta}$ values for $\boldsymbol{\pi}$, a value smaller than 1 for l_{step}, say 0.5 or smaller, may be necessary, even if this inevitably increases the number of iterations. As a reasonable starting value for $\boldsymbol{\beta}$, one could set to zero the regression coefficients corresponding to covariates, together with the global log-odds ratios intercepts, whereas the global logits intercepts have to be chosen by taking into account the inequality constraints (2.15). The variance covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $V(\hat{\boldsymbol{\beta}}) = F_P(\hat{\boldsymbol{\beta}})^{-1}$. When a *NUPPOM* is considered, the form of matrix $\mathbf{X}_i$ is $\mathbf{X}_i = \bigoplus_{k=1}^{3}\mathbf{X}_{k,i}$, where:

$$\mathbf{X}_{k,i} = \begin{pmatrix} 1 & & 0 & \boldsymbol{x}'_{i,\mathscr{S}_k} & \boldsymbol{x}'_{i,\bar{\mathscr{S}}_k} & & \mathbf{0}' \\ & \ddots & & \vdots & & \ddots & \\ 0 & & 1 & \boldsymbol{x}'_{i,\mathscr{S}_k} & \mathbf{0}' & & \boldsymbol{x}'_{i,\bar{\mathscr{S}}_k} \end{pmatrix}.$$

Thus the full design matrix is simply $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', ..., \mathbf{X}_m')'$. The *weight function* for the ith observation is defined as $\mathbf{W}_i(\boldsymbol{\beta}) = n_i\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i'}\right)\text{diag}(\boldsymbol{\pi}_i)^{-1}\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\right)$, the weight matrix is

$\mathbf{W}(\boldsymbol{\beta}) = (\mathbf{W}_1(\boldsymbol{\beta})', \mathbf{W}_2(\boldsymbol{\beta})', \ldots, \mathbf{W}_m(\boldsymbol{\beta})')'$, the *hat matrix* is $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X} + \mathbf{P})^{-1}\mathbf{X}'\mathbf{W}(\hat{\boldsymbol{\beta}})$, and the *Akaike Information Criterion* is $AIC = -2(l(\hat{\boldsymbol{\beta}}) - tr(\mathbf{H}))$.

Appendix B: penalty terms in matrix form

When (2.10) or (2.14) is used $\mathbf{P} = \mathbf{E}'\Lambda'^{1/2}\Lambda^{1/2}\mathbf{E}$, where Λ is the matrix of smoothing values, and $\mathbf{E} = \bigoplus_{k=1}^{3} \mathbf{E}_k$.

In (2.10) matrices Λ and $\mathbf{E}_k$ depend on the penalty (ridge or ARC1). Let $\boldsymbol{d} = (D_1 - 1, D_2 - 1, (D_1 - 1)(D_2 - 1))'$ a vector indexed by d_k, $k = 1,2,3$, and let $\boldsymbol{\lambda}'_{k_{|\mathscr{I}_k^0|}} = (\lambda_{k,0}, \boldsymbol{\lambda}'_{k_{|\bar{\mathscr{I}}_k|}})$ be the smoothing values vector of length $|\mathscr{I}_k^0|$, that is the cardinality of the set of variables undergone to penalization for the kth equation in (2.7). Then

$$\Lambda = \text{diag}(\boldsymbol{\lambda}'_{1,\mathscr{P}_1^0}, \boldsymbol{\lambda}'_{2,\mathscr{P}_2^0}, \boldsymbol{\lambda}'_{3,\mathscr{P}_3^0}),$$

where $\boldsymbol{\lambda}'_{k,\mathscr{P}_k^0} = (\lambda_{k,0}\mathbf{1}'_{(d_k-1)}, \mathbf{0}'_{|\mathscr{S}_k|}, \boldsymbol{\lambda}'_{k_{|\bar{\mathscr{I}}_k|}} \otimes \mathbf{1}'_{(d_k-1)})$.

ARC1 penalty term

For the ARC1 penalty, let $\mathbf{T}_k$ be the $d_k \times d_k$ upper triangular matrix of 1's. Its inverse $\mathbf{V}_k = \mathbf{T}_k^{-1}$ has 1's on the main diagonal, -1's on the first superdiagonal and 0's elsewhere. Further, let $\mathbf{V}_{k_{-1}}$ be the matrix $\mathbf{V}_k$ ignoring the last row. Then

$$\mathbf{E}_k = (\mathbf{V}_{k_{-1}}, \mathbf{0}_{(d_k-1)\times|\mathscr{S}_k|}, \mathbf{1}'_{|\bar{\mathscr{I}}_k|} \otimes \mathbf{V}_{k_{-1}}).$$

Ridge-type penalty term

Let $\mathbf{I}_{d_k}$ be the $d_k \times d_k$ identity matrix. Then

$$\mathbf{E}_k = (\mathbf{I}_{d_k}, \mathbf{0}_{d_k \times |\mathscr{S}_k|}, \mathbf{1}'_{|\bar{\mathscr{S}}_k|} \otimes \mathbf{I}_{d_k}).$$

Lasso-type penalty term

Let $\mathbf{B}_k = \text{diag}(\boldsymbol{\beta}_k^{1/2})$, where $\boldsymbol{\beta}_k$ is the parameter vector for the k-th equation, and let $\mathbf{G}_k = \mathbf{B}_k^{-1}$ be the Moore-Penrose generalized inverse matrix of $\mathbf{B}_k$. Then,

$$\boldsymbol{E}_k = (\mathbf{G}_k, \mathbf{0}_{d_k \times |\mathscr{S}_k|}, \mathbf{1}'_{|\bar{\mathscr{S}}_k|} \otimes \mathbf{G}_k).$$

ARC2 penalty term

Let $\boldsymbol{c} = (D_1 - 1, D_2 - 1, D_2 - 1, D_1 - 1)'$ a vector indexed by c_h, $h = 1, ..., 4$, and let $k = 1, 2, 3$. Then

$$\Lambda = \text{diag}(\boldsymbol{\lambda}'_{1,\mathscr{P}_1^0}, \boldsymbol{\lambda}'_{2,\mathscr{P}_2^0}, \boldsymbol{\lambda}'_{3,\mathscr{P}_3^0}, \boldsymbol{\lambda}'_{4,\mathscr{P}_3^0}),$$

where $\boldsymbol{\lambda}'_{h,\mathscr{P}_k^0} = (\lambda_{h,0}\mathbf{1}'_{(c_h-1)}, \mathbf{0}'_{|\mathscr{S}_k|}, \mathbf{1}'_{(c_h-1)} \otimes \boldsymbol{\lambda}'_{h_{|\bar{\mathscr{S}}_k|}})$.

Define $s_{h,j}$, $j \in \bar{\mathscr{S}}^0$, the order of operator $\Delta^{s_{h,j}}$, for the jth variable, also including the intercepts. Let $\mathbf{T}_h$ be the $c_h \times c_h$ upper triangular matrix of $1's$ and let $\mathbf{V}_h = \mathbf{T}_h^{-1}$. Let $\mathbf{V}^{s_{h,j}} = \prod_{h=1}^{s_{h,j}} -\mathbf{V}_h$, let $\mathbf{V}^{s_{h,j}}_{-s_{h,j}}$ be the matrix $\mathbf{V}^{s_{h,j}}$ ignoring the last $s_{h,j}$ rows and let $\mathbf{U}_{k,\bar{\mathscr{S}}^0} = (\mathbf{U}_{k,0}, \mathbf{U}_{k,\bar{\mathscr{S}}})$, where $\mathbf{U}_{k,\bar{\mathscr{S}}} = (\mathbf{U}_{k,1}, ..., \mathbf{U}_{k,|\bar{\mathscr{S}}|})$ and $\mathbf{U}_{1,j} = \mathbf{V}^{s_{1,j}}_{-s_{1,j}}$, $\mathbf{U}_{2,j} = \mathbf{V}^{s_{2,j}}_{-s_{2,j}}$ and $\mathbf{U}_{3,j} = (\mathbf{I}_{(c_3)} \otimes \mathbf{V}^{s_{3,j}}_{-s_{3,j}}) \oplus (\mathbf{V}^{s_{4,j}}_{-s_{4,j}} \otimes \mathbf{I}_{(c_4)})$. Then

$$\mathbf{E}_k = (\mathbf{U}_{k,0}, \mathbf{0}_{(d_k-1) \times |\mathscr{S}_k|}, \mathbf{U}_{k,\bar{\mathscr{S}}}).$$

The penalty term for ordering constraints

For (2.16) let $\mathbf{N} = (\mathbf{1}_n \otimes \mathbf{E})'$, being $\mathbf{E}$ defined as for ARC1, and n the sample size. Let $\Lambda = (\lambda_1 \mathbf{I}_n, \lambda_2 \mathbf{I}_n, \mathbf{0}_{n\times n})$. Then

$$\mathbf{P} = \mathbf{X}'\mathbf{N}'\Lambda'^{1/2} I(\boldsymbol{\beta}'\mathbf{X}'\mathbf{N}') I(\mathbf{N}\mathbf{X}\boldsymbol{\beta})\Lambda^{1/2}\mathbf{N}\mathbf{X},$$

where $I(\mathbf{N}\mathbf{X}\boldsymbol{\beta} \leq 0)$ is element-wise, that is $I(a_{ij} \leq 0) = 1$ if true, 0 otherwise.

References

Bartolucci, F. and Forcina, A. (2002). Extended RC association models allowing for order restrictions and marginal modeling. *Journal of the American Statistical Association*, **97(460)**, 1192–1199.

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

Burnham, K. and Anderson, D. (2000). *Model selection and inference. A practical information - Theoretic approach*. Springer-Verlag, New York.

Bustami, R., Lesaffre, E., Molenberghs, G., Loos, R., Danckaerts, M., and Vlietinck, R. (2001). Modelling bivariate ordinal responses smoothly with examples from ophthalmology and genetics. *Statistics in Medicine*, **20**, 1825–1842.

Calvaruso, V., Cammá, C., Marco, V. D., Maimone, S., Bronte, F., Enea, M., Dardanoni, V., Manousou, P., Pleguezuelo, M., Xirouchakis, E., Attanasio, M., Dusheiko, G., Burroughs, A. K., and Craxí, A. (2010). Fibrosis staging in chronic hepatitis C: analysis of

discordance between transient elastography and liver biopsy. *Journal of Viral Hepatitis*, **17(7)**, 469–474.

Castera, L., Vergniol, J., Foucher, J., Bail, B. L., Chanteloup, E., Haaser, M., Darrieth, M., Couzigou, P., and Lédinghen, V. D. (2005). Prospective comparison of transient elastography, Fibrotest, APRI and Liver Biopsy for the assessment of fibrosis in chronic hepatitis C. *Gastroenterology*, **138**, 343–350.

Colombi, R. and Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, **88(4)**, 1007–1019.

Dale, J. R. (1986). Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses. *Biometrics*, **42(4)**, 909–917.

Dardanoni, V. and Forcina, A. (1998). A unified approach to likelihood inference on stochastic orderings in a non-parametric context. *Journal of the American Statistical Association*, **93**, 1112–1123.

Desantis, S. D., Houseman, E. A., Coull, B. A., Stammer-Rachamimov, A., and Betensky, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics*, **9(2)**, 249–262.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.

Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, **50**, 61–76.

Fahrmeier, L., Gieger, C., and Heumann, C. (1999). An application of isotonic longitudinal marginal regression to monitoring the healing process. *Biometrics*, **55**, 951–956.

Gieger, C. (1997). Non- and semiparametric marginal regression models for ordinal response. Technical report, University of Munich, Institute of Statistics. URL http://epub.ub.uni-muenchen.de/.

Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, **57**, 533–546.

Goodman, L. A. (1979). Simple models for the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, **74(367)**, 537–552.

Gray, R. J. (1994). Spline-based test in Survival Analysis. *Biometrics*, **50**, 640–652.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Lapp, K., Molenberghs, G., and Lesaffre, E. (1998). Models for the association between ordinal variables. *Computational Statistics and Data Analysis*, **28**, 387–411.

Muggeo, V. M. R. and Ferrara, G. (2008). Fitting generalized linear models with unspecified link function: A p-spline approach. *Computational Statistics and Data Analysis*, **52**, 2529–2537.

Muggeo, V. M. R. and Tagliavia, M. (2010). A flexible approach to the crossing hazards problem. *Statistics in Medicine*, **29**, 1947–1957.

Peterson, B. and Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **39**(2), pp. 205–217. ISSN 00359254.

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

Qaqish, B. F. and Ivanova, A. (2006). Multivariate logistic models. *Biometrika*, **93(4)**, 1011–1017.

Tutz, G. and Scholz, T. (2003). Ordinal regression modelling between proportional odds and non-proportional odds. Technical report, University of Munich, Institute of Statistics. URL http://epub.ub.uni-muenchen.de/.

Tutz, G. (2003). Generalized semiparametrically structured ordinal models. *Biometrics*, **59** (2), 263–273. ISSN 1541-0420.