Multivariate Ordered Logit Regressions

May 13, 2005

Valentino Dardanoni, University of Palermo Antonio Forcina, University of Perugia[1]

ABSTRACT

In this paper we combine recent advances in marginal modelling for contingency tables with the notion of copula to formulate a class of models for describing how the joint distribution of a set of ordinal response variables depends on exogenous regressors. We derive the main properties of a marginal parameterization, the *global interaction copula*, whose nature is essentially non parametric, and study its relation with an underlying seemingly unrelated system of latent variables regression. An efficient algorithm for maximum likelihood estimation is described; we also derive the asymptotic distribution of the likelihood ratio under suitable equality and inequality constraints. An application to O-levels grades for a cohort of UK students clarifies the usefulness of these models.

JEL codes: C35, C12, C13.

Keywords: ordered response variables, system of latent regressions, marginal parameterization, multivariate copulas, chi-bar squared distribution.

1 Introduction

Many interesting problems in economics involve the study of how an ordered response variable depends on a set of regressors. One way to model such data is to assume that the observed response is the discrete version of a continuous latent variable for which a linear regression model holds; alternatively, an *index model* may be written for a given transformation (*link function*) of the discrete probabilities. It is well known that these two approaches (the latent and index models) are essentially equivalent (see section 2.1 below). Choice of the logistic

[1]Valentino Dardanoni (corresponding author): Department of Economics, University of Palermo, Viale delle Scienze, 90128 Palermo, Italy; vdardano@unipa.it

Antonio Forcina: Department of Statistics, University of Perugia, Via Pascoli, 06100 Perugia, Italy; forcina@stat.unipg.it.

distribution for the errors in the latent model or the logit link in the index model gives rise to the ordered logit model, while the normal distribution and the probit link gives rise to the ordered probit model. Both models are very well known and increasingly widely used.

On the other hand, often we may be concerned with modelling how *several* closely related ordered response variables depend on regressors. Related sets of response variables may arise when we consider different achievements or choices made by each subject at a given point in time, or repeated choices on the same item made by each subject at different points in time as with *panel data*. Though for panel data specific approaches have been designed to take into account individual heterogeneity and serial correlation (Honoré and Kyriazidou, 2000, Honoré and Lewbel, 2002 and Magnac, 2004), they are mainly restricted to the binary case. For more general contexts, the standard way to approach these types of problems is by assuming that the set of underlying latent variables forms a *seemingly unrelated regression system*. When the model is completed by assuming that the errors follow a multivariate normal distribution and the observed responses are discretizations of the underlying latent variables, we obtain the *multivariate ordered probit model* which, however, has been implemented only for the case of binary responses. The main limitation of this approach is that maximum likelihood estimation requires the computation of multivariate normal integrals which, in most cases, must be computed by Monte Carlo simulation (see Greene, 2004, p.714-5 for a general discussion and Cappellari and Jenkins, 2003, for the practical implementation, in the binary case, of simulated likelihood inference).

In this paper we attempt to combine recent advances in marginal modeling for contingency tables with the notion of copula (Nelsen, 1999) to formulate a class of multivariate ordered logit models with exogenous regressors and provide simple procedures for likelihood inference. Copulas provide, essentially, a non parametric way of modelling a joint distribution by combining two or more univariate distributions and is particularly appropriate for dealing with a set of response variables when these can conveniently be treated in a symmetric fashion. Recent developments on contingency tables have considered parameterizations which are not log-linear in an effort to (i) take into account the ordered nature of the variables and (ii) be able to model directly the marginal distributions of interest. Seminal contributions in these directions are those of Molenberghs and Lesaffre (1994) and Glonek and McCullagh (1995): they both considered regression models specific for ordered response variables and described how to transform the vector of probabilities containing the joint distribution into logits and higher order marginal interaction parameters which take into account the ordinal nature of the response. More recent advances of the *marginal modelling approach* are contained in Bergsma and Rudas (2002) who studied a very flexible way of parameterizing a contingency table with a suitable set of marginal and conditional interactions of log-linear type.

After reviewing the basic properties of the ordered logit model in the univariate case and the connection between the latent variable and index approaches in Section 2, in Section 3 we derive the main properties of the marginal modelling approach and its relation with the latent

variable model, through the notion of copula which in principle could be of nonparametric nature. Section 4 discusses the computation and properties of maximum likelihood estimates, and the asymptotic distribution of the likelihood ratio under suitable equality and inequality constraints. Section 5 clarifies the usefulness of these models with an application where we examine the effects of covariates on O-level scores for a cohort of UK students.

2 The univariate case

2.1 The latent logistic regression model and the ordered logit model

Suppose we want to investigate how an ordinal variable Y taking value in $\{1, \ldots, m\}$ depends on a vector of covariates $\boldsymbol{z}$. A line of approach would be to assume that the ordinal variable is the discretized version of an underlying continuous variable which depends on covariates as in a linear regression model. More precisely let

$$Y^* = \alpha + \boldsymbol{z}'\boldsymbol{\beta} + \epsilon \tag{1}$$

and assume that Y is determined by the latent Y^* and a set of additional parameters $\gamma_1, \gamma_i \ldots, \gamma_{m-1}$ known as *cut points* as follows:

$$Y \leq j \Leftrightarrow Y^* \leq \gamma_j \quad (j = 1, \ldots, m-1) \tag{2}$$

Under the additional assumption that ϵ has a standard logistic distribution with *cdf* $L(t) = P(\epsilon \leq t) = \exp(t)/[1 + \exp(t)]$ and $\boldsymbol{z}$ is exogeneous, it follows that

$$P(Y \leq j \mid \boldsymbol{z}) = P[\epsilon \leq \gamma_j - (\alpha + \boldsymbol{z}'\boldsymbol{\beta})] = \frac{\exp(\gamma_j - \alpha - \boldsymbol{z}'\boldsymbol{\beta})}{1 + \exp(\gamma_j - \alpha - \boldsymbol{z}'\boldsymbol{\beta})}$$

which implies that the *global logits* of the manifest variable (which can be seen as the natural extension of the binary logits to ordered responses) are linear in $\boldsymbol{\beta}$

$$\log[P(Y > j \mid \boldsymbol{z})/P(Y \leq j \mid \boldsymbol{z})] = \boldsymbol{z}'\boldsymbol{\beta} - (\gamma_j - \alpha). \tag{3}$$

Note that, since the intercept α and the cut points cannot be identified simultaneously, we can assume, without loss of generality, that $\alpha = 0$.

The model described above, known in the statistical literature as the *proportional odds model* (McCullagh, 1980), may also be derived as an extension of the ordinary logistic model for binary data. More precisely, if we define $m-1$ binary variables $Y^{(j)}$, $j = 1, \ldots, m-1$ in such a way that $Y^{(j)} = 1$ whenever $Y > j$ and assume that the logits of $Y^{(j)}$ have the form $\boldsymbol{z}'\boldsymbol{\beta} - \gamma_j$, we obtain exactly the same model as (3). By doing so, we have constrained the vector of regression coefficients to be the same across the binary variables $Y^{(j)}$, $j = 1, \ldots, m-1$. Global logits can thus be interpreted as binary logits computed on successive splits of the ordinal response categories into two categories.

Call *ordered logit* the model defined by (3) when formulated as an extension of a binary regression, and *latent regression* the model based on (1) and (2). The well known relationship between these two approaches is stated formally in the following:

Theorem 1 *If ϵ has a standard logistic distribution, the parameters $\boldsymbol{\beta}$, $\gamma_1, \ldots, \gamma_{m-1}$ are the same in the latent regression and in the ordered logit models.*

It is also known that these two approaches can be generalized as follows. Let $G : R \to [0, 1]$ be any continuous and strictly increasing function and assume that the error in the latent regression model has cumulative distribution G. Then the parameters $\boldsymbol{\beta}$, $\gamma_1, \ldots, \gamma_{m-1}$ are the same in the latent regression model with $\alpha = 0$ and in the $G - link$ ordered regression model. In other words, the choice of a link function $G^{-1}(P(Y \leq j))$ for the ordered model is equivalent to the choice of the cumulative distribution of ϵ in the latent regression model. The best known alternative to the ordered logit model is of course the ordered probit model, where G is the standardized normal *cdf*.

2.2 Stochastic orderings

It is interesting to note that the global logits satisfy a stochastic ordering property which seems particularly appropriate when the response variables have an ordinal nature, in the sense that their relevant properties should be preserved under arbitrary monotonic transformations. We have

Theorem 2 *Given two discrete ordered random variables Y_1 and Y_2 in $\{1, \cdots, m\}$, the following conditions are equivalent:*

1. *$E[u(Y_1)] \geq E[u(Y_2)]$ for any function $u(y)$ which is non decreasing;*
2. *Y_1 is stochastically greater than Y_2;*
3. *$\log[P(Y_1 > j)/P(Y_1 \leq j)] \geq \log[P(Y_2 > j)/P(Y_2 \leq j)]$ $\}j < m$.*

Proof The equivalence between the first two conditions is well known (see for example Hadar and Russell (1969) Theorems 1 and 2). The equivalence with the third condition follows by noting that global logits are strictly increasing transformations of the cumulative distribution. Q.E.D.

The result above implies that, if we regress a global logit on a given covariate and the regression coefficient is positive, then the response variable becomes stochastically larger whenever that covariate increases; because of this, regression coefficients in the ordered logit regression have a direct interpretation in terms of stochastic dominance. Notice instead that the standard interpretation of ordered logit coefficients (see for example Crawford, Pollak and Vella, 1998), which refers to the density rather than the cumulative distribution of the response variable, implies often a rather convoluted interpretation.

2.3 Extension to non proportional odds

Equation (3) could be generalized by allowing the regression coefficients to depend on j:

$$\log[P(Y > j \mid \boldsymbol{z})/P(Y \leq j \mid \boldsymbol{z})] = \boldsymbol{z}'\boldsymbol{\beta}_j - \gamma_j, \quad j = 1, \ldots, m-1; \tag{4}$$

this model is sometimes called the *generalized ordered logit* model (see Fu, 2004 and references therein). Notice that by choosing the first logit as reference, (4) could be written as

$$\log[P(Y > j \mid \boldsymbol{z})/P(Y \leq j \mid \boldsymbol{z})] = -\gamma_j + \boldsymbol{z}'\boldsymbol{\beta}_1 + \boldsymbol{z}'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_1) = \boldsymbol{z}'\boldsymbol{\beta} - (\gamma_j - \boldsymbol{z}'\boldsymbol{\delta}_j) \quad (j = 1, \cdots, m-1)$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_1$, $\boldsymbol{\delta}_1 = \mathbf{0}$ and $\boldsymbol{\delta}_j$ can be interpreted as the effect of certain covariates in shifting the jth threshold relative to the first. In other words, the generalized ordered logit model (4) has the following latent variable interpretation: when it is up to the subject to locate his/her response in a given category, we may assume that the thresholds are no longer constant but may be affected by individual covariates. Notice that, in the equation above, any category $h < m$ could be taken as reference category by putting $\boldsymbol{\beta}_h = \boldsymbol{\beta}$ which implies $\boldsymbol{\delta}_h = \mathbf{0}$.

Especially when m is large, the more general model (4) may require many more parameters than the simple ordered logit model. This may be justified only when it is reasonable to assume that the threshold between adjacent categories depends on subjective judgments, as for instance in the analysis of the determinants of *health status, happiness* etc. Because the ordered logit model may be seen as a properly constrained generalized logit model, the effect of covariates on cut points may be tested, as we shall see in section 4.1, by imposing appropriate linear constraints. When the dependence of cut point on individual covariates is not justified by the nature of the response variable, the rejection of the proportional odds assumption should be taken as a warning that the latent model is not properly specified, like when, for instance, the distribution of the error is heteroscedastic or $\boldsymbol{z}$ is not exogeneous.

3 The multivariate case

3.1 The multivariate latent logistic regression model

Suppose we want to relate the joint distribution of K ordinal variables $Y_1, \ldots, Y_K$, with Y_k taking values in $\{1, \ldots, m_k\}$, to a vector of covariates $\boldsymbol{z}$. In itself the latent regression equation (1) is easily extended to the multivariate case by assuming the existence of K latent continuous variables Y_k^*, $k = 1, \ldots, K$, such that $Y_k \leq j$ whenever $Y_k^* \leq \gamma_{j)k=}$, with

$$Y_k^* = \boldsymbol{g}_k(\boldsymbol{z})'\boldsymbol{\beta} + \epsilon_k, \tag{5}$$

where $\boldsymbol{g}_k(\boldsymbol{z})$, $k = 1, \ldots, K$, are known functions of $\boldsymbol{z}$ defined in such a way that each latent regression has, possibly, its own subset of regressors. A model with such a structure could be justified when there are K different ordered choices made by each unit at a given point

of time, or when each unit makes an ordered choice on a given item at K different points of time, like in a panel data model with free correlations across periods but without lagged dependent variables. Let H denote the joint distribution of the errors so that $H(a_1, \ldots, a_K) = P(\epsilon_1 \leq a_1, \ldots, \epsilon_K \leq a_K)$, and let H_k, $k = 1, \ldots, K$ denote the corresponding univariate marginal distributions. Theorem 1 then implies the existence of K separate ordered logit models

$$\log \frac{P(Y_k > j(k) \mid \boldsymbol{z})}{P(Y_k \leq j(k) \mid \boldsymbol{z})} = -\gamma_{j(k)} + \boldsymbol{g}_k(\boldsymbol{z})'\boldsymbol{\beta}, \quad (k = 1, \ldots, K,\ j = 1, \cdots, m_k - 1) \tag{6}$$

by letting each marginal distribution H_k be the standard logistic cumulative distribution. However there are a few reasons for modelling also the joint distribution of $(\epsilon_1, \ldots, \epsilon_K)$:

1. to limit our consideration to the univariate (marginal) distributions is equivalent to assume that $(\epsilon_1, \ldots, \epsilon_K)$ are independent, a mis-specification of the model which is likely to produce a loss of efficiency;

2. if we wish, we can impose restrictions on the coefficients across equations;

3. the degree of association between $(\epsilon_1, \ldots, \epsilon_K)$ may be of interest on its own, may depend on covariates and provide evidence of residual unexplained heterogeneity.

3.2 Multivariate copulas

A simple way of modelling the association among the errors in the K equations, in a way which is, possibly, non parametric and treats the component variables in a symmetric fashion, is through the notion of a *copula* which we now review briefly. A copula is a function $C : [0,1]^K \to [0,1]$ which *connects* a set of univariate distributions into a joint distribution. In particular, Sklar's Theorem implies that any continuous distribution H may be determined by its marginal distributions H_k and a copula $C_H(u_1, \ldots, u_K) = H(H_1^{-1}(u_1), \ldots, H_K^{-1}(u_K))$, $u_k \in [0,1]$ for all k; thus a copula describes the association structure of H irrespective of its univariate marginals. Nelsen (1999, Chapter 1) provides an excellent introduction to copulas and their properties. This tool is particularly appropriate for describing the association between ordinal variables because copulas are invariant to strictly monotone transformations of the random variables; the other basic property of copulas is that univariate marginals and association structure may be modelled separately and then combined. Therefore, the latent regression model (5) above can be described by the set of regression coefficients $\boldsymbol{\beta}$, the cut points $\gamma_{j(k)}$, the univariate marginal distribution of the ϵ_k and the copula C_H which in turn may also depend on covariates.

A well known family of parametric copulas is, for instance, the Gaussian copula:

$$C_{\boldsymbol{\rho}}(u_1, \ldots, u_K) = \Psi_K(\Psi_1^{-1}(u_1), \ldots, \Psi_1^{-1}(u_K), \boldsymbol{\rho})$$

where Ψ_Q denotes the standard Q-variate normal distribution and $\boldsymbol{\rho}$ is the $K(K-1)/2$ vector of all bivariate correlation coefficients. When the Gaussian copula is combined with K standard normal marginal distributions it gives the multivariate normal distribution, which, when

employed in the latent regression model (5) gives rise to the *multivariate ordered probit* model. Though this model may look appealing because it is based on the apparently simplest and most familiar distribution, it has certain drawbacks. First, when K is greater than 3, computation of cumulative probabilities, which are required by maximum likelihood estimation algorithms, becomes intractable and Monte Carlo approximations are required. Secondly, the Gaussian copula implies a specific (and rather restrictive) parametric form for the association structure among the errors.

3.3 The global interaction copula

Conditionally on $\boldsymbol{z}$ the distribution of the response variables Y_k, $k = 1, \ldots, K$ follows a multinomial distribution with $t = \prod_1^K m_k$ cells. The vector containing the cell probabilities of such a distribution will be denoted by $\boldsymbol{\pi}(\boldsymbol{z})$; its entries are ordered lexicographically by letting variables with a larger index run faster. Let Π denote the t-dimensional simplex. Though, formally, any mapping $\lambda(\boldsymbol{z}) = \lambda(\boldsymbol{\pi}(\boldsymbol{z})) : \Pi \rightarrow \Re$ may be called a *parameter*, we are interested in parameters which describe relevant aspects of the joint distribution such as the univariate marginal distributions and their association. Also, it would be desirable to work with a suitable set of parameters which provide an *invertible mapping* so that the approach is non parametric to start with, a substantial difference with respect to the Gaussian copula. However, because the number of parameters of the saturated model increases much faster than K, we propose a flexible approach within which it is natural to impose parametric restrictions which affect only the less relevant aspects of the joint distribution.

Parameters defined as contrasts among the logarithms of probabilities of belonging to disjoint subsets of cells will be called *interactions*, see e.g. Bartolucci, Colombi and Forcina (2005). The global logits may be seen as interaction parameters of the first order since they involve first order differences. They may also be called *marginal* because they are computed within the marginal distribution of the variables to which they refer. A general definition of global marginal interaction parameters is given below:

Definition 1 *A* global marginal interaction parameter *is determined by the following elements:*

1. *the set of the variables involved which can be denoted by $\mathcal{I}$, a subset of $\mathcal{Q} = (1, \ldots, K)$;*
2. *the set $\mathcal{M} \subset \mathcal{Q}$ of the variables defining the joint distribution where the interaction is defined;*
3. *the vector of indices $\boldsymbol{i}$ which, for any variable Y_k, $k \in \mathcal{I}$, denotes the level where dichotomization takes place, with $i(k) < m_k$;*
4. *the vector of indices $\boldsymbol{j}$ and a corresponding binary vector $\boldsymbol{b}$ that indicates, for each variable Y_k, $k \in (\mathcal{M}/\mathcal{I})$, that $Y_k \leq j(k)$ if $b(k) = 0$ and $Y_k > j(k)$ if $b(k) = 1$;*

Notice that when $\mathcal{I} = \mathcal{M}$, the vectors $\boldsymbol{j}$, $\boldsymbol{b}$ are empty. A given global marginal interaction parameter derived from the joint distribution $\boldsymbol{\pi}(\boldsymbol{z})$ will be denoted by $\lambda_{\mathcal{I};\mathcal{M}}(\boldsymbol{z}; \boldsymbol{i} \mid \boldsymbol{j}; \boldsymbol{b})$; because in the sequel we deal only with interaction parameters of type global, this qualification will be omitted. Bartolucci, Colombi and Forcina (2005) have shown that a marginal interaction parameter for the variables in $\mathcal{I}$ computed within $\mathcal{M}$ may be defined by the following recursive equation

$$\begin{aligned} \lambda_{k;\mathcal{M}}(\boldsymbol{z}; i(k) \mid \boldsymbol{j}; \boldsymbol{b}) &= \log[\pi_{\mathcal{M}}(\boldsymbol{z}; i(k), \boldsymbol{j}; 1, \boldsymbol{b})] - \log[\pi_{\mathcal{M}}(\boldsymbol{z}; i(k), \boldsymbol{j}; 0, \boldsymbol{b})] \quad \} \\ \lambda_{\mathcal{I}\cup k;\mathcal{M}}(\boldsymbol{z}; \boldsymbol{i}, j(k) \mid \boldsymbol{j}; \boldsymbol{b}) &= \lambda_{\mathcal{I};\mathcal{M}}(\boldsymbol{z}; \boldsymbol{i} \mid j(k), \boldsymbol{j}; \boldsymbol{b}, 1) - \lambda_{\mathcal{I};\mathcal{M}}(\boldsymbol{z}; \boldsymbol{i} \mid j(k), \boldsymbol{j}; \boldsymbol{b}, 0), \end{aligned}$$

where $\pi_{\mathcal{M}}(\boldsymbol{z}; j(k), \boldsymbol{j}; b_k, \boldsymbol{b})$ denotes the marginal probability within $\mathcal{M}$ that $Y_k \leq j(k)$ or $Y_k > j(k)$ according to whether $b_k = 0$ or 1 and similarly for all $l \in \mathcal{M}\backslash k$, $Y_l \leq j(k)$ or $Y_l > j(k)$ depending on the value of b_l. The first equation defines the logits of Y_k and the second says that the interaction between the variables in $\mathcal{I} \cup k$ can be written as the difference between the two interactions in $\mathcal{I}$ constrained, respectively, to $Y_k > j(k)$ and $Y_k \leq j(k)$. Thus, the logits introduced in section 2.1 are marginal interaction parameters of the first order. The marginal interaction parameters of the second order are the so-called *global log-odds ratios*; they depend on a pair of cut points and may be seen as contrasts between the logit of one variable conditional on the other taking a "high" value versus a "low" value. The log-odds ratios for the pair (Y_h, Y_k) are denoted as $\lambda_{(h,k);(h,k)}(\boldsymbol{z}; j(h), j(k)$, for all possible combination of $j(h) < m_h$ and $j(k) < m_k$, so that there are $(m_h - 1)(m_k - 1)$ of them, each one corresponding to a 2×2 table obtained by splitting the levels of Y_h, Y_k into "low" and "high", according to the pair of cut points. These features are further clarified by the examples below.

Example 1 *The logit of Y_1 at cut point 3 defined within $\mathcal{M} = \{1\}$ is simply*

$$\lambda_{1;1}(\boldsymbol{z}; 3) = \log[\pi_1(\boldsymbol{z}; 3; 1)] - \log[\pi_1(\boldsymbol{z}; 3; 0)].$$

The log-odds ratio between Y_1, Y_2 defined within the marginal $\mathcal{M} = \{1, 2\}$ at the cut point (3,4) may be written as

$$\lambda_{\mathcal{M};\mathcal{M}}(\boldsymbol{z}; (3, 4)) = \lambda_{1;\mathcal{M}}(\boldsymbol{z}; 3 \mid 4; 1) - \lambda_{1;\mathcal{M}}(\boldsymbol{z}; 3 \mid 4; 0).$$

Note that the recursive equation is symmetric with respect to the component variables, so we can also write

$$\lambda_{\mathcal{M};\mathcal{M}}(\boldsymbol{z}; (3, 4)) = \lambda_{2;\mathcal{M}}(\boldsymbol{z}; 4 \mid 3; 1) - \lambda_{2;\mathcal{M}}(\boldsymbol{z}; 4 \mid 3; 0).$$

In order to define the third order interaction between Y_1, Y_2, Y_3 we need to define two log-odds ratios, say, between Y_i, Y_2 where Y_3 switches between $Y_3 > j(3)$ and $Y_3 \leq j(3)$.

Let $\boldsymbol{\lambda}(\boldsymbol{z})$ be the vector of size $v \leq t - 1$ that contains all marginal interaction parameters of the joint distribution $\boldsymbol{\pi}(\boldsymbol{z})$ which are considered to be of interest, for all the possible combinations of the levels of the response variables. The elements of this vector are arranged so that

its first $\sum_1^K (m_k - 1)$ elements are the marginal logits for $Y_1, \ldots, Y_K$. Let $\mathcal{M}_1, \ldots \mathcal{M}_s$ be the list of marginals, in non decreasing order, where the interactions of interest are defined. For example, if only logits and log-odds ratios are considered (so that all higher order interactions are set equal to zero) $s = K + K(K-1)/2$. For any $r \leq s$, let $\mathcal{F}_r$ be the collection of all $\mathcal{I} \subseteq \mathcal{Q}$ such that the corresponding marginal interaction parameters are defined within any marginal up to $\mathcal{M}_r$. Then we have

Definition 2 *(Bergsma and Rudas, 2002). A vector of marginal interaction parameters* $\boldsymbol{\lambda}(\boldsymbol{z})$ *constitute a* hierarchical *parameterization if, for any interaction* $\mathcal{I}$ *which is contained in at least one of the marginals* $(\mathcal{M}_1, \ldots, \mathcal{M}_r)$, $\mathcal{I} \in \mathcal{F}_r$.

The following definition provides a simple way of handling the multinomial distribution within the exponential family of distributions:

Definition 3 *Any matrix* $\boldsymbol{K}$ *which is of full rank and such that* $\boldsymbol{K}\mathbf{1} = \mathbf{0}$ *defines a vector of* canonical parameters *for the multinomial distribution* $\boldsymbol{\theta}(\boldsymbol{z}) = \boldsymbol{K} \log[\boldsymbol{\pi}(\boldsymbol{z})]$. *A vector of canonical parameters which is the log-linear analog of* $\boldsymbol{\lambda}(\boldsymbol{z})$ *may be constructed as follows. For any* $\mathcal{I} \in \mathcal{F}_s$ *let*

$$\boldsymbol{K}_{\mathcal{I}} = \bigotimes_{k=1}^{K} \boldsymbol{K}_{\mathcal{I},k}, \} \boldsymbol{K}_{\mathcal{I},k} = \begin{cases} \boldsymbol{D}_k & \} k \in \mathcal{I} \\ \boldsymbol{u}_k & \} \end{cases},$$

where $\boldsymbol{D}_k$ *is the* $(m_k - 1) \times m_k$ *matrix of first order differences between adjacent terms and* $\boldsymbol{u}_k$ *is a vector of dimension* m_k *having 1 as first entry and 0 elsewhere. Then construct* $\boldsymbol{K}$ *by stacking the matrices* $\boldsymbol{K}_{\mathcal{I}}$ *one below the other with the same order as in* $\boldsymbol{\lambda}(\boldsymbol{z})$.

Let $\Pi_{\mathcal{F}} \subseteq \Pi$ be the set of all probability vectors such that the marginal interactions $\lambda_{\mathcal{I};\mathcal{M}}(\boldsymbol{z}; \boldsymbol{i} \mid \boldsymbol{j}; \boldsymbol{b})$ for all $\mathcal{I} \notin \mathcal{F}_s$ are constrained to zero; let also $\boldsymbol{G}$ be the left inverse of $\boldsymbol{K}$. Obviously, when $\Pi_{\mathcal{F}} = \Pi$, $v = t - 1$.

Lemma 1 *Given any hierarchical vector of marginal interactions* $\boldsymbol{\lambda}(\boldsymbol{z})$*, there exists a matrix of contrasts* $\boldsymbol{C}$ *and a matrix* $\boldsymbol{M}$ *of zeros and ones such that*

$$\boldsymbol{\lambda}(\boldsymbol{z}) = \boldsymbol{C} \log[\boldsymbol{M}\boldsymbol{\pi}(\boldsymbol{z})] \quad \} \boldsymbol{\pi}(\boldsymbol{z}) = \frac{\exp[\boldsymbol{G}\boldsymbol{\theta}(\boldsymbol{z})]}{\mathbf{1}' \exp[\boldsymbol{G}\boldsymbol{\theta}(\boldsymbol{z})]} \quad \} \boldsymbol{\theta}(\boldsymbol{z}) \in \Re^v;$$

the mapping from $\boldsymbol{\pi}(\boldsymbol{z})$ *to* $\boldsymbol{\lambda}(\boldsymbol{z})$ *is invertible and differentiable for all strictly positive* $\boldsymbol{\pi}(\boldsymbol{z}) \in \Pi_{\mathcal{F}}$.

Proof Both results can de derived from Bartolucci, Colombi e Forcina (2005); in particular see their Appendix for the first claim while the second claim is a special case of their Theorem 1. They also describe a simple algorithm for constructing the $\boldsymbol{C}$, $\boldsymbol{M}$ matrices. Q.E.D.

Remark 1 *The marginal interaction parameters are not* variation independent*, for instance, the logits must be strictly decreasing and the log-odds ratios must be such that the bivariate cumulative distribution is non decreasing. The space of compatible marginal interaction parameters is the set* $\mathcal{L} = \{\boldsymbol{\lambda}(\boldsymbol{z}) : \boldsymbol{\lambda}(\boldsymbol{z}) = \boldsymbol{C} \log[\boldsymbol{M}\boldsymbol{\pi}(\boldsymbol{z})]$ *and* $\boldsymbol{\pi}(\boldsymbol{z}) \in \Pi_{\mathcal{F}}\}$.

Lemma 2 *Let $\boldsymbol{\lambda}(\boldsymbol{z}) \in \mathcal{L}$ be a vector of marginal interactions; the mapping from $\boldsymbol{\theta}(\boldsymbol{z})$, or equivalently $\boldsymbol{\pi}(\boldsymbol{z})$, to $\boldsymbol{\lambda}(\boldsymbol{z})$ may be inverted by the following algorithm:*

1. *at the initial step choose a value $\boldsymbol{\theta}^{(0)}$such that $\boldsymbol{\lambda}^{(0)}$ is sufficiently close to $\boldsymbol{\lambda}(\boldsymbol{z})$;*
2. *at the h-th step update the vector of canonical parameters by the first order approximation*

$$\boldsymbol{\theta}^{(h)} = \boldsymbol{\theta}^{(h-1)} + \boldsymbol{D}[\boldsymbol{\lambda}(\boldsymbol{z}) - \boldsymbol{\lambda}^{(h-1)}]$$

   where $\boldsymbol{D} = \partial\boldsymbol{\theta}/\partial\boldsymbol{\lambda}' = [\boldsymbol{C}diag(\boldsymbol{M}\boldsymbol{\pi})^{-1}diag(\boldsymbol{\pi})\boldsymbol{G}]^{-1}$
3. *iterate until the norm of $\boldsymbol{\lambda}(\boldsymbol{z}) - \boldsymbol{\lambda}^{(h-1)}$ is close to 0.*

Proof This is a direct application of the Newton algorithm. Since the mapping from $\boldsymbol{\theta}(\boldsymbol{z})$ to $\boldsymbol{\lambda}(\boldsymbol{z})$ has continuous second order derivatives $\forall\, \boldsymbol{\theta}(\boldsymbol{z})$ whose elements are finite, the result may be derived, for example, from Theorem 4.4 in Suli and Mayers (2003). In our experience, by setting $\boldsymbol{\theta}^{(0)} = \mathbf{0}_{t-1}$, the algorithm always converges as long as $\boldsymbol{\lambda}(\boldsymbol{z})$ is not too close to the boundary of the parameter space; this may happen when one or more elements are much larger than 20 in modulus. Q.E.D.

Definition 4 *A multivariate ordered regression model is a system of v equations of the form*

$$\boldsymbol{\lambda}(\boldsymbol{z}) = \boldsymbol{\lambda}_0 + \boldsymbol{Z}\boldsymbol{\beta} \tag{7}$$

where $\boldsymbol{\lambda}_0$ is a vector of intercepts and the matrix $\boldsymbol{Z}$ is a matrix of known constants which depend on $\boldsymbol{z}$.

Many restrictions of interest, like assuming that some interactions do not depend on covariates and/or on cut points, can be easily imposed. We will discuss some of these restrictions in the sequel.

The key result is contained in the following theorem which essentially says that the logits determine the regression model (5) while the higher order interaction parameters define the association structure of the errors. From a technical point of view, the interaction parameters defines only a *subcopula* (see Nelsen, 1999, p. 39), which is like a copula determined on a discrete grid. The fact that the extension of a subcopula to a copula is not unique, implies that certain features of the underlying multivariate latent distribution are not identifiable; because of this it is sufficient to show that there is at least one multivariate distribution with the required characteristics.

Theorem 3 *Assume that, for each possible vector $\boldsymbol{z}$, $\boldsymbol{\lambda}(\boldsymbol{z})$ belongs to $\mathcal{L}$ and that in the multivariate latent regression equations (5) the errors $(\epsilon_1, \ldots, \epsilon_K)$ have univariate marginals which are standard logistic. Let also $C_H(u_1, \ldots, u_K \mid \boldsymbol{z}) = H(L^{-1}(u_1), \ldots, L^{-1}(u_K) \mid \boldsymbol{z})$ denote the copula corresponding to the joint distribution of the errors; then*

1. *$Y_k^* \leq \gamma_{j(k)}$ is equivalent to $Y_k \leq j(k)$ if and only if $\gamma_{j(k)}$ is the intercept of $\lambda_{k;k}(\boldsymbol{z}; j(k))$ and the block of $m_k - 1$ rows of $\boldsymbol{Z}$, which corresponds to the kth logit, are equal to $\boldsymbol{g}_k(\boldsymbol{z})'$;*

2. *the elements of $\boldsymbol{\lambda}(\boldsymbol{z})$ which correspond to the marginal interactions of order greater than one are consistent with the copula C_H which describes the joint distribution of the errors.*

PROOF. Because of the univariate logistic distribution of the errors, $Y_k^* \leq \gamma_{j(k)} \Leftrightarrow Y_k \leq j(k)$ if and only if $\lambda_{k;k}(\boldsymbol{z}; j(k)) = \boldsymbol{g}_k(\boldsymbol{z})'\boldsymbol{\beta} - \gamma_{j(k)}$; the first statement of the theorem follows from the fact that this holds for all $\boldsymbol{z}$. This also implies that

$$H(\gamma_{j(1)} - \boldsymbol{g}_1(\boldsymbol{z})'\boldsymbol{\beta}, \ldots, \gamma_{j(K)} - \boldsymbol{g}_K(\boldsymbol{z})'\boldsymbol{\beta} \mid \boldsymbol{z}) = P(Y_1 \leq j(1), \ldots, Y_K \leq j(K) \mid \boldsymbol{z});$$

from Lemmas 1 and 2 it follows that $\boldsymbol{\lambda}(\boldsymbol{z})$ determines completely the joint distribution of $(Y_1, \ldots, Y_K) \mid \boldsymbol{z}$ as well as the joint distribution of the errors $H(\epsilon_1, \ldots, \epsilon_K \mid \boldsymbol{z})$ in the corresponding finite grid of cut points. In particular, the set of marginal interaction parameters of order greater than one define the subcopula

$$C[P(Y_1 \leq j(1)), \ldots, P(Y_K \leq j(K)) \mid \boldsymbol{z}]$$

over all possible combinations of $j(1), \ldots, j(K)$; this can be extended to a copula in many ways. A simple multilinear extension may be constructed as follows. Let $q_{0k} = P(Y_k \leq j(k))$, $q_{1k} = P(Y_k \leq j(k) + 1)$ and $\tau_k = [u_k - q_{0k}]/[q_{1k} - q_{0k}]$ for all $u_k \in [q_{0k}, q_{1k}]$; let also $\boldsymbol{b}$ denote a binary vector of dimension K and $\boldsymbol{q}(\boldsymbol{b})$ the vector which has elements q_{0k} if $b(k) = 0$ and q_{1k} if $b(k) = 1$, then the following equation is a copula which is consistent with the subcopula defined above

$$C_H = \sum_{\boldsymbol{b}} C[\boldsymbol{q}(\boldsymbol{b})] \prod_1^K (\tau_k)^{b_k} (1 - \tau_k)^{1 - b_k}$$

where the sum is for all 2^K possible binary vectors $\boldsymbol{b}$. Because the continuous latent distribution is not identifiable outside of the finite grid, there is no restriction in assuming that it has this specific form. Q.E.D.

3.4 Positive quadrant dependence

The log-odds ratios, which determine the association for any pair of responses, are also closely related to the notion of *positive quadrant dependence* (PQD), an instance of positive dependence between ordinal variables first introduced by Lehmann (1966): two random variables Y_h, Y_k are PQD if

$$Pr(Y_h \leq j(h), Y_k \leq j(k)) \geq Pr(Y_h \leq j(h)) Pr(Y_k \leq j(k))\} j(h), j(k),$$

which intuitively means that, compared to the case of independence, "small" values of Y_h tend to go with "small" values of Y_k. Negative quadrant dependence is defined by reversing the inequality above.

The ordinal nature of Y_h and Y_k seems to motivate the requirement that their relevant properties should be preserved under arbitrary monotonic transformations. The following result in the theory of stochastic orderings links the notion of positive dependence, PQD, to the log-odds ratios:

Theorem 4 *Given a pair of discrete ordered random variables Y_h and Y_k, taking values respectively in $\{1, \cdots, m_h\}$ and $\{1, \cdots, m_k\}$, and any pair of increasing functions u, v, the following conditions are equivalent:*

1. *$Cov[u(Y_h), v(Y_k) \mid \boldsymbol{z}] \geq 0$;*
2. *Y_h and Y_k are PQD conditionally on $\boldsymbol{z}$;*
3. *$\lambda_{\{h,k\};\{h,k\}}(\boldsymbol{z}; j(h), j(k)) \geq 0 \; \} j(h) < m_h, j(k) < m_k$.*

Proof See Nelsen (1999), exercises 5.22 and 5.27, p. 153-4. Q.E.D.

This result may be interpreted as saying that, if the ordered variables are the discrete version of continuous latent variables discretized at arbitrary cut points, the log-odds ratios are the most appropriate measure of association, in the sense that the sign of the dependence between the underlying variables is preserved, irrespective of how the ordered categories are constructed. This can be compared with the correlation coefficient, which is the association parameter in the multivariate ordered probit model.

4 Statistical inference

4.1 Hypotheses of interest

A convenient feature of the multivariate ordered regression model defined by equation (7) is that many relevant hypotheses can be expressed in the form of linear equality and inequality constraints on $\boldsymbol{\beta}$. For example, when the dimension of the K-way table is reasonably large, it may be convenient to constrain to zero higher order interactions to avoid working in a too large parameters space; by imposing such restrictions, very little is lost because the higher is the order of an interactions, the less direct is its interpretation. However, though this hypothesis could be expressed as a linear constraint on $\boldsymbol{\beta}$, it is computationally more efficient if such parameters are not included in $\boldsymbol{\lambda}(\boldsymbol{z})$, so that its size v is strictly smaller than $t - 1$. The most restrictive strategy would be to set equal to zero all interactions of order greater than two; in this case $v = \sum_1^K (m_k - 1) + \sum_{h=1}^{K-1} \sum_{k=h+1}^{K} (m_h - 1)(m_k - 1)$.

A more flexible approach would be to assume that higher order interaction parameters are constant irrespective of the cut points, an assumption which is the multivariate analog of the *Plackett distribution*. The family of bivariate Plackett distributions, introduced by Plackett (1965), has been extended to the multivariate case by Molenberg and Lesaffre (1994). A model where all interactions of order grearter than two are set equal to zero, and the Plackett

restriction is imposed by means of linear equality constraints on $\boldsymbol{\beta}$, provides a close analog to the multivariate ordered probit model, with the correlation coefficients replaced by the corresponding bivariate log-odds ratios. One advantage of the Plackett copula is that, because the log-odds ratios may assume any real value, they may be easily linked to covariates by a regression model, while the correlation coefficients in the probit model are usually assumed to be unaffected by covariates.

The simplest example of equality constraints arise when one or more regression coefficients in the latent regressions are assumed to be zero. The proportional odds model may be seen as a model where the regression coefficients are constrained to be equal across different cut points. Linear inequality constraints may be used to test a stochastic dominance effect of certain covariates on a set of latent regressions. We could also be interested to test positive dependence between a pair of responses against conditional independence by imposing that all the $(m_h - 1)(m_k - 1)$ log-odds ratios are positive against being zero; alternatively, we could test that a given pair of responses has a stronger positive dependence than an other pair if the variables involved have the same number m of levels, by constraining the corresponding log-odds ratios to be greater.

In the sequel, let

$$\mathcal{H} : \{\boldsymbol{\beta} : \boldsymbol{E\beta} = \mathbf{0},\, \boldsymbol{U\beta} \geq \mathbf{0}\}$$

denote an hypothesis specified by appropriate choice of the "equality" and "inequality" matrices $\boldsymbol{E}$ and $\boldsymbol{U}$, where $(\boldsymbol{E}^{'}, \boldsymbol{U}^{'})'$ has full rank; clearly, the case with $\boldsymbol{E}$ or $\boldsymbol{U}$ equal to the null matrices correspond to restriction with only inequalities or only equalities respectively.

4.2 Likelihood inference

Suppose we have independent observations $(Y_{1i}, \ldots, Y_{Ki}, \boldsymbol{z}_i)$ for a sample of n units. Let $\boldsymbol{t}(\boldsymbol{z}_i)$ be a vector of size $\prod m_k$, with value 1 for the element corresponding to the observed combination of $(Y_1, \ldots, Y_K)$ for the ith unit, and 0 otherwise. To simplify notations, in the sequel we write $\boldsymbol{t}(i)$ instead of $\boldsymbol{t}(\boldsymbol{z}_i)$; a similar convention will be adopted for any vector which depends on $\boldsymbol{z}_i$. Under independent sampling, conditionally on $\boldsymbol{z}_i$, $\boldsymbol{t}(i)$ has a multinomial distribution with vector of probabilities $\boldsymbol{\pi}(i)$. In order to manipulate the likelihood function, we write the multinomial as an exponential family with the same vector of canonical parameters used in Definition 3; in practice, these are all the log-linear interactions $\mathcal{I}$ which are in $\mathcal{F}_s$, so that $\boldsymbol{\lambda}(i)$ has the same dimension as $\boldsymbol{\theta}(i)$.

The contribution of the ith unit to the log likelihood is

$$L(i) = \boldsymbol{t}(i)' log[\boldsymbol{\pi}(i)]$$

which, by expressing $\boldsymbol{\pi}(i)$ in terms of the canonical parameters, may be written as

$$L(i) = \boldsymbol{t}(i)' \boldsymbol{G\theta}(i) - log[\mathbf{1}' exp(\boldsymbol{G\theta}(i))].$$

Recall that

$$\boldsymbol{D}(i) = \frac{\partial \boldsymbol{\theta}(i)}{\partial \boldsymbol{\lambda}(i)'} = [\boldsymbol{C} diag[\boldsymbol{M}\boldsymbol{\pi}(i)]^{-1}\boldsymbol{M} diag[\boldsymbol{\pi}(i)]\boldsymbol{G}]^{-1}.$$

The chain rule and simple calculations give the individual score vector:

$$\boldsymbol{s}(i) = \frac{\partial L(i)}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\lambda}(i)'}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\theta}(i)'}{\partial \boldsymbol{\lambda}(i)} \frac{\partial L(i)}{\partial \boldsymbol{\theta}(i)} = \boldsymbol{Z}(i)'\boldsymbol{D}(i)'\boldsymbol{G}'[\boldsymbol{t}(i) - \boldsymbol{\pi}(i)],$$

and the individual contribution to the expected information matrix:

$$\boldsymbol{F}(i) = -E\left[\frac{\partial^2 L(i)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right] = \boldsymbol{Z}(i)'\boldsymbol{D}(i)'\boldsymbol{G}'\boldsymbol{\Omega}(i)\boldsymbol{G}\boldsymbol{D}(i)\boldsymbol{Z}(i)$$

where $\boldsymbol{\Omega}(i) = diag[\boldsymbol{\pi}(i)) - \boldsymbol{\pi}(i)\boldsymbol{\pi}(i)']$ is the kernel for the variance matrix of the multinomial distributions.

Having assumed that the units are independent, the log-likelihood is $L(\boldsymbol{\beta}) = \sum_i L(i)$, thus the score vector $\boldsymbol{s}(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ and the expected information matrix $\boldsymbol{F}(\boldsymbol{\beta}) = -E(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'})$ can be obtained by summing over units.

4.3 Parameter estimation

Maximum likelihood estimates of $\boldsymbol{\beta}$ under $\mathcal{H}$ can be obtained by an algorithm which extends to inequality constraints the seminal algorithm introduced by Aitchison and Silvey (1959). Alternatively, the constrained method of scoring could be used as in Bartolucci and Forcina (2000), but even though it is usually faster, it works only as long as the updated estimate produces compatible values of $\boldsymbol{\lambda}(i)$ for each subject. The Aitchison and Silvey algorithm is based on iterated linear approximations of the regression model onto the space of the canonical parameters which are variation independent; the approximation is updated until convergence. Formally:

- assign a starting value $\boldsymbol{\beta}^{(0)}$ which produce compatible $\boldsymbol{\lambda}(i)$ for all units;
- at the hth step, compute a linear approximation of $\boldsymbol{\theta}(i)$ and a quadratic approximation of the log likelihood

$$\begin{aligned} \boldsymbol{\theta}(i)^h &= \boldsymbol{\theta}(i)^{h-1} - \boldsymbol{D}(i)^{h-1}\boldsymbol{\lambda}(i)^{h-1} + \boldsymbol{D}(i)^{h-1}\boldsymbol{Z}(i)\boldsymbol{\beta}^{h-1} \\ Q_h(\boldsymbol{\beta}) &= -(\boldsymbol{\beta} - b^h)'\boldsymbol{F}^{h-1}(\boldsymbol{\beta} - b^h)/2 \end{aligned}$$

  where $\boldsymbol{b}^h = [\boldsymbol{F}^{h-1}]^{-1}[\boldsymbol{s}^{h-1} + \boldsymbol{s}_1^{h-1}]$ and $\boldsymbol{s}_1^{h-1} = \sum_i \boldsymbol{Z}(i)'\boldsymbol{D}(i)^{h-1\prime}\boldsymbol{G}'\boldsymbol{\Omega}(i)^{h-1}\boldsymbol{G}\boldsymbol{D}(i)^{h-1}\boldsymbol{\lambda}(i)^{h-1}$
- set $\boldsymbol{\beta}^h$ to be equal to the constrained maximum of $Q_h(\boldsymbol{\beta})$ under $\mathcal{H}$,
- iterate until convergence.

The starting point must be chosen so that the corresponding $\boldsymbol{\lambda}(i)^0$ is compatible for all subjects. This may be achieved by setting to zero the intercepts of association parameters and all the regression coefficients corresponding to the covariates. Notice that when inequality constraints are present, so that $\boldsymbol{U}$ is not the null matrix, the maximization of $Q_h(\boldsymbol{\beta})$ at each step requires a quadratic optimization which is itself iterative; there are many algorithms for quadratic optimization under inequality constraints, which are usually very fast and reliable.

Theorem 5 *If the A-S algorithm converges, it converges to a local maximum of $L(\boldsymbol{\beta})$ subject to $\mathcal{H}$.*

PROOF It is easily verified that both the likelihood function and the transformation from $\boldsymbol{\theta}(i)$ to $\boldsymbol{\lambda}(i)$ satisfy the conditions discussed by Aitchison and Slvey (1959) p. 817; it follows that, as n increases, the probability that a constrained maximum exists tends to one. If the algorithm converges, it must converge to a local maximum by the argument of Aitchison and Slvey (1959), p. 826. Q.E.D.

Notice that our parameterization satisfies the two basic assumptions given by Rao (1973, p.296), namely identifiability and continuity of the transformation from $\boldsymbol{\beta}$ to $\boldsymbol{\pi}$. It follows that, provided that $\boldsymbol{\beta}_0$, the true value of $\boldsymbol{\beta}$ under $\mathcal{H}$, is an interior point of the parameter space, then the *m.l.e.* of $\boldsymbol{\beta}$ under $\mathcal{H}$ exists and is consistent.

4.4 Hypotheses testing

Assume that we have selected a *reference model* with parameters vector $\boldsymbol{\beta}$: this can be taken as the model such that all the hypotheses of interest may be expressed by suitable sets of equality and/or inequality constraints on $\boldsymbol{\beta}$. Let $\dot{\boldsymbol{\beta}}$ denote the unrestricted *m.l.e* of $\boldsymbol{\beta}$ in the reference model, $\hat{\boldsymbol{\beta}}$ the *m.l.e.* of $\boldsymbol{\beta}$ under $\mathcal{H}$ and $\bar{\boldsymbol{\beta}}$ the *m.l.e.* of $\boldsymbol{\beta}$ under $\mathcal{H}_0 = \{\boldsymbol{\beta} : \boldsymbol{E\beta} = \mathbf{0},\ \boldsymbol{U\beta} = \mathbf{0}\}$.

¿From standard asymptotic results it follows that, if $\boldsymbol{Z}(i)$ is of full rank for all i and, as n increases, $\boldsymbol{F}(\boldsymbol{\beta})/n$ remains of full rank, $\dot{\boldsymbol{\beta}}$ has an asymptotic normal distribution $N(\boldsymbol{\beta}, \boldsymbol{F}(\boldsymbol{\beta})^{-1})$. Therefore, hypotheses on single elements of $\boldsymbol{\beta}$ may be tested by comparing the estimate with the corresponding standard error. Joint testing may be based on the asymptotic distribution of the LR statistic; recall the well known result that the LR for testing the reference model against $\mathcal{H}_0$

$$T_{02} = 2(L(\dot{\boldsymbol{\beta}}) - L(\bar{\boldsymbol{\beta}}))$$

has asymptotic χ^2_r distribution where r is the number of linearly independent columns in $(\ \boldsymbol{E}'\quad \boldsymbol{U}'\ )$.

When inequalities are involved, because $\mathcal{H}_0$ does not coincide with $\mathcal{H}$, the testing problem may be split into testing the reference model against $\mathcal{H}$ and testing $\mathcal{H}$ against $\mathcal{H}_0$. The corresponding LR statistics may be written as

$$\begin{aligned} T_{12} &= 2(L(\dot{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}})) \\ T_{01} &= 2(L(\hat{\boldsymbol{\beta}}) - L(\bar{\boldsymbol{\beta}})). \end{aligned}$$

It is also useful to recall the follwing:

Definition 5 *Let $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{V})$ be a k-dimensional normal random vector, and let $\mathcal{C}$ be a polyhedral cone in R^k. The chi-bar-squared random variable $\bar{\chi}^2(\mathcal{C}, \boldsymbol{V})$ is equal to:*

$$\bar{\chi}^2(\mathcal{C}, \boldsymbol{V}) = \boldsymbol{b}'\boldsymbol{V}^{-1}\boldsymbol{b} - min_{\boldsymbol{a}\in\mathcal{C}}(\boldsymbol{b}-\boldsymbol{a})'\boldsymbol{V}^{-1}(\boldsymbol{b}-\boldsymbol{a})$$

and has distribution function:

$$Pr(\bar{\chi}^2(\mathcal{C}, \boldsymbol{V}) \leq x) = \sum_{i=0}^{k} w_i(\mathcal{C}, \boldsymbol{V}) F_\chi(x, i)$$

where $F_\chi(x, i)$ denotes the distribution function of a chi-square with i d.f. and $w_i(\mathcal{C}, \boldsymbol{V})$ is the probability that the projection of $\boldsymbol{b}$ onto $\mathcal{C}$ belongs to a face of dimension i.

We have:

Theorem 6 *Under the assumption that the true value $\boldsymbol{\beta}_0$ belongs to the interior of $\mathcal{H}_0$, the asymptotic distributions of T_{01} and T_{12} are:*

$$T_{01} \rightarrow \bar{\chi}^2(\mathcal{C}_\mathcal{H}, \boldsymbol{F}^{-1}(\boldsymbol{\beta}_0))$$

$$T_{12} \rightarrow \bar{\chi}^2(\mathcal{C}^0_\mathcal{H}, \boldsymbol{F}^{-1}(\boldsymbol{\beta}_0))$$

where $\mathcal{C}_\mathcal{H}$ denotes the convex cone defined by $\mathcal{H}$ and $\mathcal{C}^0_\mathcal{H}$ is its dual in the metric determined by the information matrix at $\boldsymbol{\beta}_0$.

Proof We give only a sketch of the proof, which can be derived, for example, from Shapiro (1985). Let $\tilde{\boldsymbol{\beta}}$ denote any of the ML estimates of $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma} = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\boldsymbol{x} = F(\boldsymbol{\beta}_0)^{-1}\boldsymbol{s}(\boldsymbol{\beta}_0)/\sqrt{n}$. Since $\tilde{\boldsymbol{\beta}}$ is consistent, a second order expansion of $L(\tilde{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}_0$ may be written as

$$L(\tilde{\boldsymbol{\beta}}) = L(\boldsymbol{\beta}_0) + \boldsymbol{s}'(\boldsymbol{\beta}_0)\boldsymbol{F}(\boldsymbol{\beta}_0)^{-1}\boldsymbol{s}(\boldsymbol{\beta}_0)/(2n) - (\boldsymbol{x}-\boldsymbol{\gamma})'\boldsymbol{F}(\boldsymbol{\beta}_0)(\boldsymbol{x}-\boldsymbol{\gamma})/2 + o_p(\mid \boldsymbol{\gamma} \mid^2).$$

This is a special case of equation (3.4) in Andrews (1999) so that the last term can be replaced by an $o_p(1)$ term. It follows that T_{12} and T_{01} are asymptotically equivalent, respectively, to:

$$min_{\boldsymbol{\gamma}\in\mathcal{H}}(\boldsymbol{x}-\boldsymbol{\gamma})'\boldsymbol{F}(\boldsymbol{\beta}_0)(\boldsymbol{x}-\boldsymbol{\gamma})$$

and

$$min_{\boldsymbol{\gamma}\in\mathcal{H}_0}(\boldsymbol{x}-\boldsymbol{\gamma})'\boldsymbol{F}(\boldsymbol{\beta}_0)(\boldsymbol{x}-\boldsymbol{\gamma}) - min_{\boldsymbol{\gamma}\in\mathcal{H}}(\boldsymbol{x}-\boldsymbol{\gamma})'\boldsymbol{F}(\boldsymbol{\beta}_0)(\boldsymbol{x}-\boldsymbol{\gamma})$$

and the result follows because $\boldsymbol{x}$ has asymptotic normal distribution $N[\boldsymbol{0}, \boldsymbol{F}(\boldsymbol{\beta}_0)^{-1}]$. Q.E.D.

Asymptotic critical values for these statistics depend on the probability weights $w_i(\mathcal{C}_\mathcal{H}, \boldsymbol{F}^{-1}(\boldsymbol{\beta}_0))$ which are difficult to compute; reliable estimates may be obtained by Monte Carlo simulation as described by Dardanoni and Forcina (1998), who also give examples of bounds for these kinds of distributions. Since $\mathcal{H}$ is a composite hypothesis, one should search for the value of $\boldsymbol{\beta} \in \mathcal{H}$ which gives the least favorable asymptotic null distribution for T_{12} and, as Wolak (1991) has shown, this value does not necessarily belong to $\mathcal{H}_0$. Dardanoni and Forcina (1998) discuss some practical solutions to this problem. Finally, notice that the joint distribution of T_{01} and T_{12} can also be derived; see Dardanoni and Forcina (1999) for details, where use of this joint distribution for hypotheses testing is also compared with alternative testing procedures.

5 An application to O-levels grades in the UK

In the British educational system, students at the age of 16 take the so called O-levels exams on a set of chosen topics. Typically each student takes a number of exams from 3 up to a dozen, depending on whether (s)he wants to access higher education. The results of the exams are letter grades. We use data from the National Child Development Survey (NCDS). This data set is a UK cohort study targeting all the population born in the UK between the 3rd to the 9th of March 1958. Individuals were surveyed at different stages of their life and information on their schooling results and their background was collected. We restrict attention to 4 O-level subjects: English, Mathematics, Art and History, because they were the most popular choices in the sample. We selected all students who took exams on all of these four subjects; the resulting sample is made of 1888 students. Letter grades are converted into 3 categories, good, fair and poor. Table 1 shows the marginal frequencies in the sample

Table 1: Marginal frequencies

	English	Art	History	Math
poor	385	561	707	779
fair	525	494	440	462
good	978	833	741	647

¿From the available data we constructed a set of 10 covariates that are listed in Table 2 below which indicates also the percentage of missing values. Within each covariate we first replaced missing values with the average of the observed values and then constructed a dummy which equals one if the value of the corresponding covariate is missing. When two covariates were missing almost always simultaneously, for the sake of parsimony, we constructed a single dummy taking value 1 when both covariates were missing. We combined education and age of the parents into the corresponding average and constructed a single missing indicator taking value 0, 1 or 2 depending on the number of missing values; the resulting variable was treated as a numerical covariate. Sex was coded with 0 for males and 1 for females and abilities at 7 and 11 have been centered at 0. Finally, for the sake of parsimony we also treated father's social class as a numerical variable giving scores 1 for low, 2 for middle and 3 for high class.

Table 2: List of covariates

Covariate	Average	short name	type
Sex	0.58	Sex	qualitative
Math ability at 7	0	M7	dummy
Read ability at 7	0	R7	numerical
Math ability at 11	0	M11	numerical
Read ability at 11	0	R11	numerical
Social class of father	1.23	Scf	numerical
Average number of years of education of parents	15.63	Ped	numerical
Average age of parents at birth	29.73	Pag	numerical
Average number of students in class at 7 and 11	34.70	Csz	numerical
Measure of parents' interest in the student's career	3.16	Pin	numerical
M7 and R7 missing	0.11	A7m	dummy
M11 and R11 missing	0.13	A11m	dummy
Scf missing	0.14	Scm	dummy
Ped missing	0.54	Pem	numerical
Pag missing	0.15	Pam	numerical
Csz missing	0.15	Csm	dummy
Pin missing	0.56	Pim	dummy

In order to find a suitable model, we used the following procedure. The starting model was the multivariate latent logistic model (5), with all 17 covariates affecting each latent regression, and the association structure of the errors such that log-odds ratios are unrestricted bur independent of covariates and all higher order interactions constrained to zero. This model requires 100 parameters: 8 intercepts for logits, 68 = 4x17 regression coefficients and 24 = 4x6 log-odds ratios. This model was initially tested against:

1. the larger model where the 32 = 8x4 *three-way interactions* are unconstrained; because $T_{02} = 37.58$ with a p-value of 0.229, it seems reasonable to assume that interactions of order 3 are indeed 0;

2. the more restrictive model that the log-odds ratios are constant as in a Plackett distribution; because $T_{02} = 35.54$ with 18 dof and a p-value of 0.008, this restriction has to be rejected.

The next step was to investigate whether the generalized ordered logit model holds for the marginal distributions. When all covariates are allowed to have different slopes (recall equation 4), T_{02} is equal to 84.36 with 68 dof and a p-value of 0.0869. Since in addition all the z-ratios were smaller than one, we decided to retain the assumption of proportional odds for all covariates, this in spite of the fact that, if taken one at a time, it was possible to find a few covariates with significantly different slopes for some single response. However, because these

occurrences did not seem to follow any meaningful pattern, we believe that they are a product of sample fluctuation. At this stage we removed from the latent regression the dummy missing indicators for social class of father (Scm), number of students in class (Csm) and parents' age (Pam) because all the corresponding z-ratios where non significantly different from zero; because removing these covariates gives $T_{02} = 8.9$ with 12 dof, this model will be the starting point for further investigations.

We next allowed log-odds ratios to depend on covariates. To have a meaningful interpretation of the intercepts and the slopes of the log-odds ratios, we centered to zero all covariates. The model with log-odds ratios depending on all the covariates was rejected as a whole; however a few interesting significant effects detected through the z-ratios were investigated further. In particular, some variables referring to family background were found significant for several pairs of responses, possibly implying that unexplained heterogeneity may depend on family background. Thus, we ended up allowing parent's interest and education to affect all the bivariate associations. Compared with the previous case where the copula does not depend on covariates, we get a $T_{02} = 21.8$, with 12 dof and a p-value of 0.0398. This model may be considered our final model. Its most relevant parameters are displayed in Table 3 and 4, which give the estimated regression coefficients and corresponding z-ratios.

Estimated values indicate that female do much better than males in English and Art, only slightly better in History and much worse in Maths. Early ability test scores seem in general very important predictors of future performance, with abilities at 11 being most important. Parents' education and parents' interest seem to have a substantial positive effect on the performance in all subjects while father social class and age of parents have a more limited and less consistent effect. The fact that, apart from History, class size does not have the expected negative effect, may be due to confounding: larger classes may be associated with better teachers or more stimulating surroundings. The fact that the coefficients of the missing indicator for Pin are all negative, indicate that probably parents' interest (among those where the value is missing) was smaller on average. On the other hand, it looks as if students with missing ability or parent's education have, on average, a larger value of the same covariates.

The fact that the regression coefficient for parental interest on the log-odds ratios are negative may be explained by assuming that individual heterogeneity is smaller among individuals whose parents' showed more concern in the education of their children. The fact that the effect of parents' education on the six log-odds ratios cannot be ignored gives a $T_{02} = 12.96$ with 6 dof and a p value of 0.0437; the one sided hypothesis that these coefficient are all less or equal 0 gives a T_{01} of 12.34 with a p value of 0.0124. This is an example of the gain in power which may be obtained by testing one-sided rather than generic alternatives.

Table 3: Estimated coefficients and z-ratios (in parentheses) for the marginal distributions

	ENGLISH	ART	HISTORY	MATH
cut point 1	1.8997	1.0877	0.6109	0.4600
	(25.9909)	(18.7734)	(11.6755)	(8.1301)
cut point 2	0.0362	-0.3492	-0.5558	-1.0192
	(0.6671)	(-6.6988)	(-10.7452)	(-16.9204)
Sex	0.8552	0.8977	0.1631	-0.4745
	(8.5091)	(9.4298)	(1.7681)	(-4.8185)
M7	0.1758	0.1004	0.0715	0.2699
	(3.0822)	(1.8564)	(1.3312)	(4.6465)
R7	0.3034	0.2099	0.0778	0.1312
	(4.8967)	(3.4363)	(1.2931)	(1.8727)
M11	0.4126	0.4202	0.5035	1.2521
	(6.2268)	(6.6268)	(7.8967)	(16.4636)
R11	0.8424	0.5205	0.4000	0.1424
	(11.9611)	(8.1812)	(6.4268)	(2.1675)
Scf	0.1190	0.0782	0.1728	0.1760
	(1.4559)	(1.0096)	(2.2484)	(2.1268)
Ped	0.2609	0.2123	0.1987	0.1952
	(6.6457)	(6.0091)	(5.8046)	(5.4599)
Pag	0.0066	0.0098	0.0047	0.0261
	(0.7070)	(1.1056)	(0.5376)	(2.8215)
Csz	0.0075	-0.0164	-0.0057	0.0071
	(1.0997)	(-2.5378)	(-0.8974)	(1.0510)
Pin	0.1776	0.2799	0.1895	0.1717
	(2.3163)	(3.8521)	(2.6185)	(2.1969)
A7m	0.9788	0.6842	0.2338	0.2555
	(3.1204)	(2.2900)	(0.7968)	(0.8177)
A11m	0.5750	0.5485	0.1896	-0.0645
	(1.9945)	(1.9881)	(0.6993)	(-0.2243)
Pedm	0.1364	0.0933	0.0830	0.0718
	(2.4285)	(1.7763)	(1.6125)	(1.3166)
Pinm	-0.3977	-0.1634	-0.0200	-0.0830
	(-2.9055)	(-1.2531)	(-0.1559)	(-0.6035)

Table 4: Estimated coefficients and z-ratios (in parentheses) for the copula

	ExA	ExH	ExM	AxH	AxM	HxM
GOR(1,1)	1.9011	1.4916	0.9950	1.7946	1.1889	1.3662
	(12.3230)	(9.8754)	(6.1304)	(13.9278)	(8.7474)	(11.0056)
GOR(1,2)	2.0117	1.5141	1.0442	1.7569	0.8619	1.1463
	(9.0681)	(7.5813)	(5.0402)	(11.0115)	(5.6555)	(8.1844)
GOR(2,1)	1.3172	1.1730	0.8712	1.6393	0.8979	1.2356
	(9.6722)	(9.7295)	(6.8756)	(12.7736)	(7.1825)	(9.5580)
GOR(2,2)	1.3008	1.3621	1.0992	1.7420	0.8427	1.1207
	(11.0074)	(11.2735)	(8.2643)	(14.8729)	(6.8276)	(9.3167)
Ped	0.0570	-0.0442	-0.1152	-0.1561	0.0025	-0.0627
	(0.7878)	(-0.6113)	(-1.5068)	(-2.3323)	(0.0359)	(-0.9463)
Pin	-0.4538	-0.1783	-0.2012	-0.2271	-0.2869	-0.1746
	(-3.0517)	(-1.2081)	(-1.2614)	(-1.5953)	(-1.9119)	(-1.1882)

6 References

Aitchison, J. and Silvey, S. D. (1958): "Maximum Likelihood Estimation of Parameters Subject to Restraints", *Annals of Mathematical Statistics*, vol. 29, pp. 813-828.

Andrews, D. W. K. (1999): "Estimation When a Parameter is on a Boundary", *Econometrica*, vol. 67, pp. 1341-1383.

Bartolucci, F. and Forcina, A. (2000): "A Likelihood Ratio Test for MTP_2 With Binary Variables", *Annals of Statistics*, vol. 28, pp. 1206-18.

Bartolucci, F., Colombi, R. and Forcina, A. (2005): "An Extended Class of Marginal Link Functions for Modeling Contingency Tables by Equality and Inequality Constraints", *Technical Report*.

Bergsma, W. P. and Rudas, T. (2002): "Marginal Models for Categorical Data", *Annals of Statistics*, vol. 30, pp. 140-159.

Cappellari, L. and Jenkins S. (2003): "Multivariate Probit Regression Using Simulated Maximum Likelihood", *The Stata Journal*, vol. 3, pp. 221-222.

Crawford, D. L., Pollack, R.A. and Vella, F. (1998): "Simple Inference in Multinomial and Ordered Logit", *Econometric Review*, vol. 17 pp. 28999.

Dardanoni, V. and Forcina, A. (1998): "A Unified Approach to Likelihood Inference on Stochastic Orderings in a Nonparametric Context," *Journal of the American Statistical Association*, 93, 1112–23.

Dardanoni, V. and Forcina, A. (1999): "Inference for Lorenz Curve Orderings," *Econometrics Journal*, vol. 2, 49-75.

Fu, V.K. (2004): "Module sg88: Gologit", *STATA manual*, STATA Corporation.

Glonek, G. J. N. and McCullagh, P. (1995): "Multivariate Logistic Models", *Journal of the Royal Statistical Society B*, vol. 57, pp. 533-546.

Greene, W.H. (2004): *Econometric Analysis*, Prentice Hall.

Hadar, J. and Russell, W. (1969): "Rules for Ordering Uncertain Prospects", *American Economic Review*, v.59, pp.25-34.

Honoré, B.E. and Lewbel, A. (2002): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors", *Econometrica*, 70, pp. 2053-2063.

Honoré, B.E. and Kyriazidou, E. (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables", *Econometrica*, 68, pp. 839-847.

Lehmann, E. L., (1966): "Some Concepts of Dependence," *The Annals of Mathematical Statistics*, 37, 1137–53.

Magnac, T. (2004): "Panel Binary Variables and Sufficiency: Generalizing Conditional Logit", *Econometrica*, 72, pp. 1859-1876.

McCullagh, P. (1980): "Regression Models for Ordinal Data", *Journal of the Royal Statistical Society B*, vol. 42, pp. 109-142.

Molenberghs, G. and Lesaffre, E. (1994): "Marginal Modeling of Correlated Ordinal Data Using a Multivariate Plackett Distribution", *Journal of the American Statistical Association*, vol. 89, pp. 633-644.

Nelsen, R.B. (1999): *An Introduction to Copulas*, Springer.

Plackett, R.L (1965): "A Class of Bivarlate Distribution," *Journal of the American Statistical Association*, Vol. 60, pp. 516-522

Rao, C. R. (1973): *Linear Statistical Inference and Its Applications*, 2nd Edition, John Wiley and Sons, New York.

Shapiro, A. (1985): "Asymptotic Distribution of Test Statistics in the Analysis of Moment Structures under Inequality Constraints," *Biometrika*, 72, 133–44.

Suli, E. and Mayers, D. (2003): *An Introduction to Numerical Analysis*, Cambridge University Press.

Wolak, F. (1991): "The Local Nature of Hypotheses Tests Involving Inequality Constraints", *Econometrica*, 59, pp. 981-95.