Modeling Repeated Ordered Categorical Data using Copulas

François Vandenhende[1,2] and Philippe Lambert[2]
1: Statistics and Information Sciences, Eli Lilly & Company,
Rue Granbonpré 11, B-1348 Mont-Saint-Guibert, Belgium,
Phone: +32-10476469, Fax: +32-10476475, *email:* francois@lilly.com.
2: Institute of Statistics, Université Catholique de Louvain,
Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.

October 27, 2000

Abstract

A parametric model for longitudinal ordered categorical data is proposed. The marginal distributions are modeled using standard regression models and the association between pairs of successive margins is characterized by a set of bivariate copula distributions under a first order Markov hypothesis. Several bivariate copula families, including Frank's, Plackett's, Clayton's, Joe's and Mardia's are used to model the association between successive responses. The corresponding log-likelihood is written in terms of the initial margin and of the transition probabilities. Parameters from the marginal regression model and from the pairwise association copula are estimated by direct likelihood maximization. The method is illustrated by analyzing data from a trial on acute migraine.

1 INTRODUCTION

The study of interest is a double-blind clinical trial involving 39 subjects suffering from acute migraine. Subjects with moderate to severe headache were randomized to three treatment groups (5 or 20 mg of LY334370 intraveneous (i.v.) or a matching placebo) and pain severity was recorded on a 4 grade scale (none, mild, moderate, severe) at 8 occasions after dosing: 0.5, 1, 1.5, 2, 3, 4, 6 and 24h. Our main interest is in the marginal groups comparison of pain scores at any time point. The dependence between successive scores recorded on the same patient is usually treated as a nuisance factor. However, the effect of treatment or time on the dependence could create interest.

After two hours of treatment, subjects were allowed to take rescue medication and to drop out of the study. Collection of severity scores was then stopped immediately but the reason for study drop-out was unfortunately not recorded.

A distinction between types of missing data is discussed in Little and Rubin [1] and models for different drop-out mechanisms are given by Diggle and Kenward [2]. In our case, missing scores are assumed to depend only on the previously observed values of a subject, but not on the subject unobserved score at drop-out. With data missing at random (MAR), the likelihood function associated to the effect parameters can be written on the observed scores while ignoring drop-outs. The investigation of a treatment effect on rescue medication use and the analysis of the dependence between response and drop-out are ongoing and will be the subject of another publication.

In this repeated measurements trial, the ordinal response is recorded together with a set of covariates on individuals across periods. Since multiple observations are taken on the same subjects, repeated responses are likely to be correlated and the dependence between these observations has to be taken into account for a proper inference. Several approaches have been used to analyze such repeated ordered categorical data. In this paper, we focus on marginal models in which the marginal probability distribution of the response is modeled as a function of covariates and time. A full likelihood method will be developed.

Compared to the generalized estimating equations method of Liang and Zeger [3], likelihood methods retain their consistency under MAR mechanisms. They also allow modeling of the dependence and formal model comparison. To the contrary, they are more sensitive to assumptions for the marginal distributions and their association. Moreover, a correct specification is required to ensure consistency. A review of full likelihood methods for multivariate categorical data is given in Molenberghs and Lesaffre [4]. Of interest is the method of Dale [5] for bivariate ordinal data. Dale [5] applies cumulative regression models to the marginal probabilities and expresses the dependence between pairs of margins in a global cross-ratio on the bivariate distributions. It generates a series of bivariate Plackett distributions [6] describing the dependence at each level of the outcomes. An extension to the multivariate case is given by Molenberghs and Lesaffre [7].

We propose to model the association between successive responses using "dependence functions", also called copulas [8], which link the corresponding marginal distributions. A copula is a continuous multivariate distribution function with uniform margins on [0,1]. Sklar [8] demonstrated that any joint distribution $H(y_1, ..., y_n)$ with marginals $F_1(y_1), ..., F_n(y_n)$ could be written as $H(y_1, ..., y_n) = C[F_1(y_1), ..., F_y(y_n)]$, where C is a copula. When $y_1, ..., y_n$ are continuous, C is unique. For discrete data, this construction is only uniquely determined on the range of the margins. Conversely, any copula function of given margins generates a multivariate distribution having these margins. Copula functions are used to model the dependence between marginal distributions. When the global cross-ratio does not depend on the level of the outcomes, the corresponding copula is the bivariate Plackett distribution [6]. Other parametric families of copula have been defined, including one or more parameters, to express a wide range of positive or negative dependence. Some one-parameter copula families will be considered in our example. A review of parametric copula families and of their properties can be found in Joe [9].

Copula models have essentially been used to model continuous multivariate survival responses. Oakes [10] and Shih and Louis [11] discussed parametric and non-parametric estimation of multivariate survival distributions. Genest and Rivest [12] suggested a non-parametric method for estimating bivariate *Archimedian* copulas, again for continuous responses. Meester and MacKay [13] applied copula models to categorical data. They used Frank's copula [14] to model symmetric dependence in clustered bivariate categorical data. The same likelihood-based method was recently applied by Trégouet et al. [15] to the analysis of clustered binary data. Using a one-parameter multivariate extension of Frank's family, the form of the generated joint distribution is identical, whatever the cluster size. The method could therefore be similarly applied to symmetric within-subject association in (possibly unbalanced) longitudinal studies.

In this paper, dependence over time is modeled as a first-order Markov process. A likelihood method is presented to estimate the marginal and dependence parameters as a function of the covariates and to compare adjustment of various models for the above mentioned data.

The layout is as follows. In Section 2, we briefly review marginal cumulative regression models and construct joint cumulative distributions using copulas. Some copula families are presented in Section 3. The relationship between copula models and the population version of Kendall's tau is established in Section 4. In Section 5, we construct the likelihood for the Markov process. The migraine data are analyzed in Section 6. Section 7 is devoted to a discussion of the method and of potential extensions.

2 MARGINAL MODEL AND JOINT DISTRIBUTION

Using standard notations, let $\{Y_{ij}, j = 1, ..., n_i\}$ be the vector of n_i repeated ordered categorical response variables with K possible values ($k = 1, ..., K$) recorded on subject i ($i = 1, ..., N$) in a trial. To conform with the example, the length n_i of the vector might be different between subjects because of the censoring of responses when rescue medications are taken. Let also assume that a set of covariates x_{ij} are recorded simultaneously. The marginal distribution at each time t_j is modeled using a generalized linear regression on the covariates, with parameter vector $\boldsymbol{\beta}$ for the covariates and α_k for the response levels. The model has the form :

$$g[\Pr(Y_{ij} \leq k)] = \alpha_k + \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{x}_{ij}, \quad \text{for} \quad k = 1, ..., K-1, \tag{1}$$

where $g(\cdot)$ is a link function.

This model applies simultaneously to all the $K-1$ cumulative probabilities and it assumes an identical effect of the covariates at each level of the distribution. A useful interpretation of (1) is that the ordinal response Y_{ij} can be expressed as a continuous latent variable Z_{ij} so that $Y_{ij} = k$ when $Z_{ij} \in (\alpha_{k-1}, \alpha_k]$, with $\alpha_0 = -\infty$ and $\alpha_K = +\infty$. The latent variable accounts for the ordering of the categories in such a way that the distribution $\Pr(Y_{ij} \leq k)$ is equal to $\Pr(Z_{ij} \leq \alpha_k)$ for $k = 1, ..., K-1$. The choice made for the link function $g(\cdot)$ drives the definition of the distribution for the latent variable. For example, when the link function is the logit, probit or complementary log-log link, the distribution becomes logistic, normal or extreme-value, respectively. More properties of the cumulative regression models are discussed in Agresti [16].

Following Sklar [8], a joint distribution for the n_i ordinal responses recorded on patient i can be created with the univariate margins from (1), using a n_i-variate copula C. The joint distribution for the observations made on subject i is then

$$\Pr(Y_{i1} \leq k_1, ..., Y_{in_i} \leq k_{n_i}) = C[\Pr(Y_{i1} \leq k_1), ..., \Pr(Y_{in_i} \leq k_{n_i})]. \tag{2}$$

3 SOME COPULA FAMILIES

In this section, we propose a review of the copula families that we shall use to analyze our data.

An important and particular copula is the product copula Π. With $U_1, ..., U_n$, a series of random variables with uniform margins on $[0, 1]$, it is defined as :

$$\Pi(u_1, ..., u_n) = \Pi_{j=1}^{n} u_j. \tag{3}$$

It generates a joint distribution with independent margins. Other multivariate copulas can also be obtained from simple bivariate families. We first start by reviewing some one-parameter bivariate copulas and discuss some higher-order extensions whenever possible.

Five one-parameter bivariate copula families are given in Table 1. These are the families of Frank [14], Plackett [6], Clayton [17], Joe [18] and Mardia [19]. The selected copulas are symmetric with respect to the margins order, C(u,v)=C(v,u), and the parameter θ models the association between the two margins. The product copula (3) is either a limiting or a particular case of each family.

Random variables U and V are positively dependent when large (or small) values for U and V are more likely to occur simultaneously than if U and V were independent. A copula distribution $C(\cdot)$ induces positive dependence if it is such that

$$C(u, v) \geq \Pi(u, v), \forall u, v \in [0, 1]. \tag{4}$$

Complete positive dependence is achieved when one variable is almost surely an increasing function of the other. This yields the copula of maximum dependence $M(u, v) = \min(u, v)$, called the *upper Fréchet bound*. Negative dependence is the reversed case, i.e. when

$$C(u, v) \leq \Pi(u, v), \forall u, v \in [0, 1]. \tag{5}$$

The copula of complete, maximum negative dependence is called the *lower Fréchet bound* and is defined as $W(u, v) = \max(u + v - 1, 0)$.

The two Fréchet inequalities (4) and (5) suggest that parametric copula families might be ordered. Positive ordering is when $C_{\theta_1}(u, v) \leq C_{\theta_2}(u, v), \forall u, v \in [0, 1]$ when $\theta_1 \leq \theta_2$. Families of Frank, Plackett, Joe and Clayton are positively ordered and the parameter θ in these families gives the sign and measures the strength of dependence between the random variables. As a counter example, the family of Mardia is not ordered. The impact of this lack of ordering on the association between responses will be further discussed in the next section.

All copula families of Table 1 allow a full range of positive dependence from Π up to M. Negative dependence is not permitted for the copula of Joe [18] but families of Frank, Plackett, Clayton and Mardia allow a full range of negative dependence down to W.

The families of Frank, Clayton and Joe are *Archimedian* [12]. This means that each of these copulas can be created from a particular function ϕ as

$$C(u, v) = \phi^{-1}[\phi(u) + \phi(v)].$$

The generating function $\phi : [0, 1] \to [0, +\infty)$ is continuous and strictly decreasing such that $\phi(0) = +\infty$ and $\phi(1) = 0$.

As proven by Kimberling [20], the function C^n from $[0, 1]^n$ to [0,1], defined as

$$C^n(u_1, ..., u_n) = \phi^{-1}[\phi(u_1) + ... + \phi(u_n)], \tag{6}$$

is an n-copula for all $n \geq 2$ if and only if ϕ^{-1} is completely monotonic on $[0, \infty)$.

One parameter n-variate copulas can be derived from the Archimedian bivariate families of Frank, Clayton and Joe using (6). However, the range for the association parameter θ is then limited to positive dependence only (from Π to M), because ϕ^{-1} is not completely monotonic under negative dependence. The extension of Clayton's copula is in Nelsen [21], p. 122 and the extension of Joe's copula is in Joe [9] p. 156. As it will be used in the example, we mention here the n-variate symmetric dependence extension of Frank's copula:

$$C_\theta(u_1, ..., u_n) = -\frac{1}{\theta}[1 + \frac{\prod_{i=1}^{n}(e^{-\theta u_i} - 1)}{(e^{-\theta} - 1)^{n-1}}]. \tag{7}$$

There are some known extensions of these families with more than one parameter, in order to capture several types of dependence between margins or to allow for negative association. Multivariate copulas with different types of dependence are discussed in Joe [9], pp. 155-165. An n-variate extension of Plackett is given by Molenberghs and Lesaffre [7], although they did not present the distribution as a copula. Since the Fréchet lower bound W^n is not a cumulative distribution function when $n > 2$, extensions of Mardia's family involving W^n are not copula functions anymore.

4 COPULA AND RANK-BASED ASSOCIATION FOR DISCRETE DATA

Rank-based association measures, such as Kendall's tau or Spearman's rho, are well-known measures of dependence between random variables. An important property of these parameters, as detailed by Hoeffding [22], is that they are "scale-invariant": they remain unchanged under strictly increasing transformations of the random variables. With categorical responses, several extensions of these measures have been introduced to account for ties in the ranking (see, e.g. Kruskal [23]). The numerical values of these measures defined for cross-classification tables depends on the grid of the table. Using simulations, Agresti [25] showed that Kendall's tau-b [24] tends to be less sensitive to grid selection than other measures. When (Y_{11}, Y_{12}) and (Y_{21}, Y_{22}) are independent and identically distributed vectors of responses, the population version of Kendall's tau-b is defined as

$$\frac{\Pr[(Y_{11} - Y_{21})(Y_{12} - Y_{22}) > 0] - \Pr[(Y_{11} - Y_{21})(Y_{12} - Y_{22}) < 0]}{\sqrt{\Pr(Y_{11} \neq Y_{21}) \Pr(Y_{12} \neq Y_{22})}}.$$

It measures the difference between the probabilities of concordance and discordance of the vectors. The denominator is involved to correct the statistic in the presence of ties. With continuous responses, it is equal to one and the formula reduces to the standard definition of Kendall's tau.

When Y_1 and Y_2 are continuous random variables, with marginal distributions F_1 and F_2 respectively, and joint distribution $C(F_1, F_2)$, the population version of Kendall's tau is given by :

$$\tau_{Y_1,Y_2} = \tau_C = 4 \int\int_{[0,1]^2} C(u,v) dC(u,v) - 1. \tag{8}$$

Being scale-invariant, the coefficient τ_{Y_1,Y_2} for the random variables Y_1 and Y_2 is the same as τ_C, the coefficient for $F_1(Y_1)$ and $F_2(Y_2)$. It is a direct function of the copula unrelated to the original marginal distributions.

For the copula functions of Table 1, Kendall's tau is a strictly increasing function of the dependence parameter θ. The double integral in (8) needs to be evaluated in order to obtain that function. A closed form for τ_C is available for the families of Clayton ($\frac{\theta}{\theta+2}$) and of Mardia ($\frac{\theta^3(2+\theta^2)}{3}$) but the integral does not have an analytical solution for the other families. Numerical approximations have to be used to establish the relationship between τ_C and θ in these cases.

With non-continuous (discrete) responses, the univariate margins are not a strictly increasing function of the random variables anymore. For instance, when no observation is made at one level of a categorical response, the cumulative probability function at the considered level is identical to that at the previous level. Although the response changes from one category to the next, its distribution remains constant. Therefore, τ_C is not necessarily equal to Kendall's tau-b, the adjusted coefficient for responses with ties (also denoted by τ_{Y_1,Y_2} in subsequent paragraphs). When Y_1 and Y_2 are discrete ordinal responses with marginal distributions $F_1(k) = \Pr(Y_1 \leq k)$, for $k = 1, ..., K$ and $F_2(l) = \Pr(Y_2 \leq l)$, for $l = 1, ..., L$ and bivariate joint distribution $C[F_1(k), F_2(l)]$, τ_{Y_1,Y_2} for Y_1 and Y_2 is calculated as

$$\frac{4\mathrm{tr}(\mathbf{C}\mathbf{\Delta_C}^{\mathrm{T}}) + \mathbf{U}^{\mathrm{T}}\mathbf{U} + \mathbf{V}^{\mathrm{T}}\mathbf{V} - \mathrm{tr}(\mathbf{\Delta_C}\mathbf{\Delta_C}^{\mathrm{T}}) - 1}{\sqrt{[\mathbf{U}^T(\mathbf{I_K} - \mathbf{U})][\mathbf{V}^T(\mathbf{I_L} - \mathbf{V})]}}, \tag{9}$$

where, with the convention that $F_1(0) = F_2(0) = 0$,

- $\mathbf{C}$ is the $(K \times L)$ matrix with $\{k, l\}^{th}$ element equal to $C[F_1(k-1), F_2(l-1)]$,
- $\mathbf{\Delta_C}$ is the $(K \times L)$ matrix with $\{k, l\}^{th}$ element equal to $C[F_1(k), F_2(l)] + C[F_1(k-1), F_2(l-1)] - C[F_1(k), F_2(l-1)] - C[F_1(k-1), F_2(l)]$,
- $\mathbf{I}_K$ is the K-dimensional unit vector,
- $\mathbf{I}_L$ is the L-dimensional unit vector,
- $\mathbf{U}$ is the K-dimensional vector with k^{th} element equal to $F_1(k) - F_1(k-1)$ and
- $\mathbf{V}$ is the L-dimensional vector with l^{th} element equal to $F_2(l) - F_2(l-1)$.

A proof of this result is in the appendix.

For independent responses, the product copula (3) is used in (9) and the equation simplifies to zero, as expected. In contrast with the continuous case, the adjusted coefficient for the discrete response is now different from the rank-based measure τ_C for the marginal distributions. It is a function not only of the copula but also of the marginal distributions. The relationship between Kendall's tau-b and the dependence parameter θ is further investigated for the copulas from Table 1.

We have considered two ordinal responses (Y_1, Y_2) both distributed across K categories, with K equal either to 4 as in the migraine example or to 10. The considered marginal distributions for the two responses are

$$\Pr[Y_j \leq k] = (k/K)^{\epsilon_j}$$

with $\epsilon_j > 0$. A discrete uniform distribution is generated for Y_j when $\epsilon_j = 1$. When $\epsilon_j < 1$ ($\epsilon_j > 1$), the probability masses are concentrated in the smallest (largest) categories.

Figures 1 and 2 present the relationship between tau-b and the parameter θ of Frank's copula for two such ordinal responses. The impact of an increase of the number of categories on the relationship is presented in Figure 1, considering uniform distributions with $K \in \{4, 10\}$. The relationship between θ and τ_C is shown on the same graph. The impact of a change of marginal distributions is detailed in Figure 2, with three sets of distributions over $K = 4$ categories each.

As Frank's family is positively ordered, Kendall's tau-b is a strictly increasing function of θ, whatever the marginal distributions. With two uniform discrete distributions (Figure 1), Kendall's tau-b increases in a similar way as τ_C over the whole range of θ. The relationship between tau-b and θ changes with the number of categories K. As K increases (from 4 to 10), distributions behave more as "continuous" distributions and we observe that tau-b tends to be closer to τ_C, although already very closed with a few categories. When marginal distributions leave uniformity (see Figure 2), the relationship also changes. It is still strictly monotone but, as marginal distributions are not uniform, the relationship between tau-b and θ is not symmetrical around tau-b=0. With non-uniform but identical distributions for the two responses (e.g. , $\epsilon = (0.25, 0.25)$), the relationship for positive dependence is preserved. However, the range for negative association is bounded below, so it will not be possible to model important negative dependences with Frank's copula when marginal distributions are similar across responses. When marginal distributions are very different and "complementary" (e.g. , when $\epsilon = (0.25, 4)$), the relationship on negative association is preserved but the range of positive association is now bounded above.

Figure 3 displays the relationship between tau-b and θ for the other copulas from Table 1. Conclusions similar to the Frank's copula can be drawn for the copulas of Plackett, Joe and Clayton since they are all positively ordered. In the previous section, we saw that Mardia's copula was not ordered. Figure 4 is a subset of Figure 3A, with tau-b $\in [-0.02, 0.02]$. Under conditions of a weak dependence, the relationship between tau-b and θ is not always strictly monotone (e.g. with $\epsilon = (0.25, 0.25)$ or $\epsilon = (0.25, 4)$). However, as illustrated in Figure 3A), when tau-b is larger in absolute value (—tau-b—¿0.02), we failed to observe a lack of monotonicity between θ and tau-b, suggesting that Mardia's copula might still be a good candidate to model moderate-to-large dependencies between categorical responses.

These results provide some insight on the relevance of using copula functions to model dependence between discrete responses in practice. In general, the simulation did not provide significant arguments against their use as the parameter θ can be interpreted as a monotone function of Kendall's tau-b. Though ordered copulas are theoretically more appealing because they preserve dependence ordering whatever the margins, limitations to the use of Mardia's copula were found very minimal and only applied to the estimation of a weak dependence. A limitation to the use of copula functions for categorical responses has been identified in two particular cases. One is when the two marginal distributions are similar but not uniform with a large negative relationship between responses. The other case is just the opposite, this is when the two marginal distributions are "complementary" but not uniform and when a large positive relationship between responses is present. In these two cases, the use of copula models might not be appropriate to quantify the dependence because the range of dependence allowed by the copula is bounded.

5 DERIVATION OF THE LIKELIHOOD

5.1 Exchangeable dependence

We first generalize the method of Meester and MacKay [13] to possibly unbalanced series of ordinal responses in a longitudinal study. Model (1) is used for the marginal distribution at each time point and an exchangeable dependence structure is modeled using the multivariate extension of Frank's copula given in (7). With these models, the contribution to the likelihood for any patient i observed up to time t_{n_i} is

$$\sum_{l_1=y_{i1}-1}^{y_{i1}} \cdots \sum_{l_{n_i}=y_{in_i}-1}^{y_{in_i}} (-1)^{\sum_j (y_{ij}-l_j)} \Pr_{\theta}(Y_{i1} \leq l_1, ..., Y_{in_i} \leq l_{n_i}). \tag{10}$$

Both the marginal distributions and the dependence parameter θ can depend on the covariates and time.

In the literature, the relationship between dependence parameters and covariates is usually not considered and dependence is treated as a nuisance factor [3]. However, when treatment effects are observed on the margins, a more effective treatment is also likely to improve the consistency of responses over time, and so, to increase their dependence. The hypothesis of an heterogeneous dependence structure across treatments will be tested in the migraine example.

5.2 Autoregressive dependence

A symmetric dependence structure assumes a constant association between responses over time. Alternative structures, which allow the dependence to decrease with the time lag separating two responses, could also be envisaged. We apply an autoregressive structure to the repeated categorical data using a first-order Markov model.

A stochastic process is a first-order Markov process when the conditional distribution of the response at any time t_j on the past history only depends on the last state of the process, i.e. when

$$\Pr(Y_{ij} \leq y_{ij} | Y_{il} = y_{il}, l = 1, ...j-1) = \Pr(Y_{ij} \leq y_{ij} | Y_{i,j-1} = y_{i,j-1}).$$

The conditional distribution $\Pr(Y_{ij} \leq y_{ij} | Y_{i,j-1} = y_{i,j-1})$ of Y_{ij} given $Y_{i,j-1}$ can be rewritten as a function of the bivariate copula C and of the marginal distributions of Y_{ij} and $Y_{i,j-1}$. It is equal to

$$\frac{C[\Pr(Y_{i,j-1} \leq y_{i,j-1}), \Pr(Y_{ij} \leq y_{ij})] - C[\Pr(Y_{i,j-1} \leq y_{i,j-1}-1), \Pr(Y_{ij} \leq y_{ij})]}{\Pr(Y_{i,j-1} \leq y_{i,j-1}) - \Pr(Y_{i,j-1} \leq y_{i,j-1}-1)}. \tag{11}$$

The contribution to the likelihood for subject i is thus

$$\Pr(Y_{i1} = y_{i1}) \prod_{j=2}^{n_i} \Pr(Y_{ij} = y_{ij} | Y_{i,j-1} = y_{i,j-1}),$$

where
$\Pr(Y_{ij} = y_{ij}|Y_{ij-1} = y_{ij-1}) =$
$\quad \Pr(Y_{ij} \leq y_{ij}|Y_{ij-1} = y_{ij-1}) - \Pr(Y_{ij} \leq y_{ij} - 1|Y_{ij-1} = y_{ij-1}).$

Using (11), we see that it is a function of the marginal models as defined in (1) and of the bivariate copula (chosen for instance in Table 1) which models the dependence between responses at successive times.

This representation allows a flexible modeling of the dependence structure. For instance, different copula families could be used to model the dependence at each time. Alternatively, a unique family with time- or covariate-varying parameter θ could also be fitted. When the same copula family with a time-independent parameter is used for each transition, a stationary process is assumed. Both stationary and non-stationary processes will be fitted to our migraine data.

5.3 Dependence between non succesive responses

We complete this section by examining the dependence induced between non-consecutive responses in the discrete first-order process. For continuous responses, Darsow, Nguyen and Olsen [26] showed how the joint bivariate distribution of the responses Y_j and Y_{j+2} could be computed as a function of the copulas for the intermediate successive responses $C(u_j, u_{j+1})$ and $C(u_{j+1}, u_{j+2})$. The bivariate joint distribution of $U_j = F(y_j)$ and $U_{j+2} = F(y_{j+2})$ is

$$\int_0^1 \frac{\partial C(u_j, t)}{\partial t} \frac{\partial C(t, u_{j+2})}{\partial t} dt. \tag{12}$$

The functional form of this new copula is generally different from that of the copula for successive responses and the dependence is likely to decrease with lags in time. A particular case is worth mentioning. If the bivariate distributions for the two pairs of successive responses are both the copula of Mardia with parameter θ, the bivariate distribution (12) is also Mardia with parameter θ^2.

With non-continuous responses, discrete derivations are required and a sum replaces the integral in (12) to compute the bivariate distribution at non-consecutive times. As a result, the form of the bivariate distribution at lagged times is more complex and does not only depend on the copulas. It also depends on the marginal distributions at the intermediate times. With ordinal responses, the bivariate distribution $P(Y_j \leq k, Y_{j+2} \leq m)$ for Y_j and Y_{j+2} is

$$\sum_l \frac{\{C[F_j(k), F_{j+1}(l)] - C[F_j(k), F_{j+1}(l-1)]\}}{F_{j+1}(l) - F_{j+1}(l-1)} \{C[F_{j+1}(l), F_{j+2}(m)] - C[F_{j+1}(l-1), F_{j+2}(m)]\}. \tag{13}$$

The bivariate distribution of responses separated by more than two measurement times (i.e. Y_j and Y_{j+k}, where $k > 2$) can be calculated using (13) iteratively.

6 ILLUSTRATION IN A MIGRAINE TRIAL

The considered trial includes 39 migraine sufferers who were treated with an investigational drug (LY334370) at two i.v. dose levels (5 mg or 20 mg), or with a matching placebo. Post-treatment pain scores over time are summarized in Table 2. Twelve, 13 and 14 patients were randomized in the placebo, 5 mg and 20 mg groups respectively, all with a moderate to severe baseline pain score. Six, 2 and 2 patients dropped out of the study before the 24-hour assessment in the placebo, 5 mg and 20 mg groups respectively. Baseline scores were comparable between groups. Kendall's tau-b coefficients are in Table 3. Responses are positively dependent. The association is decreasing with lag time and the level of association is comparable for pairs of times with comparable lag times. This suggests that a stationary Markov model could fit reasonably well.

Marginal cumulative regression models were fitted with the logit, probit and complementary log-log links. The selected model included parameters for the 3 thresholds (α_k), the log-time and a treatment indicator.

Several dependence structures were considered using the following copulas: the independence, an exchangeable structure using the n-variate extension of Frank's copula and a first order serial dependence using copulas from Table 1. Stationary and homogeneous processes were modeled using a single parameter θ across times and treatment groups. Non stationary processes were also considered when modeling serial dependence. In our case, a different dependence parameter θ was estimated at each time. Lastly, for both the exchangeable and stationary autoregressive models, the potential for an heterogeneous dependence across treatments was also investigated by fitting a separate parameter θ for each of the three groups. All parameters were jointly estimated using direct likelihood maximization and all calculations were performed using the Mathematica$^{\mathrm{TM}}$ package.

The log-likelihood for the homogeneous and stationary models is presented in Table 4 for various copula and link functions. The complementary log-log link performs slightly better whatever the dependence structure considered, except using Joe's copula. As expected from the descriptive study, the adjustment of the autoregressive model is better than that of the independence and of the exchangeable models. The complementary log-log model with serial dependence using Frank's copula provides the best fit (log lik=-188.0 with 7 parameters). The same model with a different dependence parameter for each treatment was also fitted (log lik=-186.4 with 9 parameters), without providing a significantly superior fit. Estimates of the dependence parameter θ from that model are equal to 13.5, 26.9 and 10.8 for the placebo, 5 mg and 20 mg groups respectively. Though not significant in this particular small sample study (N=39), the larger estimate of dependence in the 5 mg group ($\widehat{\theta} = 26.9$) is worth mentioning. It suggests an improved consistency in individual's responses across time and it is probably related to the smaller incidence of pain relapses observed in that group (see Table 2). Finally, a non stationary serial dependence model was also fitted based on Frank's copula with a different dependence parameter at each pair of successive times (log-lik=-184.3 with 13 parameters). The adjustment of that last model was not significantly superior to that of the homogeneous stationary model and no trend was found in parameter estimates to motivate further investigations of non-stationary processes.

Table 5 shows the parameter estimates for the selected model. Using these estimates and Equations (9) and (13), the estimates of tau-b are calculated per group and over time. They are reported in Table 6. Predicted values averaged across groups are comparable to the empirical estimates of Table 3. This indicates the good adjustment of the selected model to the data. Mean predicted tau-b for successive responses is 0.71 (s.d. = 0.09), which supports a strong positive dependence. The estimate of τ_C from (8) using Frank's copula is approximately equal to 0.75. These values are comparable to those obtained from the empirical estimates reported in bold in Table 3 (mean = 0.74, s.d. = 0.10).

As discussed in Section 4, model-based tau-b coefficients for discrete data depend on the margins. This is again illustrated in this example. Predicted tau-b (see Table 6) matrices are comparable between the 5 mg and the 20 mg group because the estimates of treatment effects (see Table 5) are similar for the two groups. To the contrary, tau-b estimates are somewhat smaller in the treated groups than under placebo and the discrepancy increases as the time interval between measures increases (see for instance values in italic in Table 6). This change is related to the significant time and treatment effects on the marginal distributions.

Importantly for the drug, the analysis has also demonstrated the effectiveness of the compound in reducing pain at either doses (5 and 20 mg), when compared to placebo. The selected model accounted for a different parameter for the 5 mg and the 20 mg groups but the estimates of these effects were not significantly different from one another. Doses of 5 mg and 20 mg have thus a similar effect in reducing pain severity at any time. However, the mode of action of the drug at these two doses might be different. We have mentioned the apparent improved consistency of scores over time when subjects are given a 5 mg dose. This suggests that the lower dose level could have a smoother and sustained effect, whereas the effect of the 20 mg dose could be stronger but also more local in time. This assumption would need to be verified in a larger scale clinical trial.

As a closing remark, parameter estimates from the marginal model were not sensitive to changes in the dependence structure (result not shown). This permits to demonstrate the robustness of our conclusions with respect to the treatment effect, but also illustrates the interest of copula-based models for discrete responses in practice.

7 DISCUSSION

A likelihood method was presented to model repeated ordered categorical data. Marginal distributions were modeled with standard cumulative regression models and the dependence between repeated responses was analyzed using copula functions. The use of copula functions allows a separate specification of the marginal and dependence models. Transitional distributions are easily calculated from the copulas and margins under a Markov hypothesis. Although only first-order models with one-parameter bivariate copulas were used, this method can be extended to higher order dependence models involving n-variate copula functions. Extension to multi-parameter copula families is straightforward. In practice, this could increase the flexibility of the modeling of the dependence in n-dimensions.

Copula models for continuous responses have the nice property that rank-based association coefficients, such as Kendall's tau, can be derived from the copula function only, independently of the marginal model. Parameters from the copula are thus directly interpretable as association measures. With discrete responses, this is not the case anymore. Estimates of the copula parameters and of the marginal model are dependent, and therefore, copula parameters cannot be strictly interpreted as a dependence measures. In Section 4, we established the relationship between Kendall's tau-b coefficient (adjusted for ties) and the copula-based distribution for categorical responses. The sensitivity of this relationship to changes in the marginal distributions was investigated in a simulation study. Results are encouraging to envisage modeling of multivariate ordinal data based on copulas. With the positively ordered copulas considered in this paper, tau-b is strictly increasing with θ so that the copula parameter θ can reasonably be interpreted as a dependence measure. This was yet verified in the migraine study in which the predicted tau-b from the autoregressive model are very close to the empirical estimates from Table 3. In most cases, the relationship between tau-b and θ is not bounded and the complete range of dependence (tau-b $\in [-1, 1]$) can be estimated using copulas. The exception to this rule is when marginal distributions are not uniform but similar (complementary) and when a strong negative (positive) dependence is observed between the two responses. In this situation, the range for tau-b is bounded below (above) and copula models might not be suitable to quantify dependence over the bounds. As a potential future research, the theoretical validation of these findings could be envisaged.

APPENDIX: POPULATION VERSION OF KENDALL'S TAU-B FOR ORDINAL DATA

The probability of concordance can be rewritten as

$$\Pr[(Y_{11}-Y_{21})(Y_{12}-Y_{22}) > 0] = \Pr(Y_{11} < Y_{21}, Y_{12} < Y_{22}) + \Pr(Y_{11} > Y_{21}, Y_{12} > Y_{22}).$$

The first term of this equation is expressed as

$$\begin{aligned}
\Pr(Y_{11} < Y_{21}, Y_{12} < Y_{22}) &= \sum_{k,l} \Pr(Y_{11} < k, Y_{12} < l) \Pr(Y_{21} = k, Y_{22} = l) \\
&= \sum_{k,l} \Pr(Y_{11} \le k-1, Y_{12} \le l-1) \Pr(Y_{21} = k, Y_{22} = l).
\end{aligned}$$

The bivariate distribution and probability mass functions can then be written in terms of the copula using (2) and (10).

Since $\Pr(Y_{11} > Y_{21}, Y_{12} > Y_{22}) = \Pr(Y_{21} < Y_{11}, Y_{22} < Y_{12})$, the same derivation can be made on the second term of the concordance part.

The probability of discordance $\Pr[(Y_{11} - Y_{21})(Y_{12} - Y_{22}) < 0]$ is rewritten as

$$\begin{aligned} &1 - \Pr(Y_{11} > Y_{21}, Y_{12} > Y_{22}) - \Pr(Y_{11} < Y_{21}, Y_{12} < Y_{22}) \\ &- \Pr(Y_{11} = Y_{21}) - \Pr(Y_{12} = Y_{22}) + \Pr(Y_{11} = Y_{21}, Y_{12} = Y_{22}). \end{aligned}$$

Observing that

$$\Pr(Y_{11} = Y_{21}) = \sum_k \Pr(Y_{11} = k) \Pr(Y_{21} = k)$$

and that

$$\Pr(Y_{11} \neq Y_{21}) = \sum_k \Pr(Y_{11} = k)[1 - \Pr(Y_{21} = k)]$$

allows to write tau-b as in (9). This completes the proof.

References

[1] Little RJA and Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New-York, 1987.

[2] Diggle PJ and Kenward MG. Informative drop-out in longitudinal data analysis. *Applied Statistics* 1994; **43**: 49–93.

[3] Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.

[4] Molenberghs G and Lesaffre E. Marginal modeling of multivariate categorical data. *Statistics in Medicine* 1999; **18**: 2237–2255.

[5] Dale JR. Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* 1986; **42**: 909–917.

[6] Plackett RL. A class of bivariate distributions. *Journal of the American Statistical Association* 1965; **60**: 516–522.

[7] Molenberghs G and Lesaffre E. Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* 1994; **89**: 633–644.

[8] Sklar A. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris.* 1959; **8**: 229–231.

[9] Joe H. *Multivariate Models and Dependence Concepts*. Chapman & Hall: London, 1997.

[10] Oakes D. Multivariate survival distributions. *Nonparametric Statistics* 1994; **3**: 343–354.

[11] Shih JH and Louis TA. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* 1995; **51**: 1384–1399.

[12] Genest C and Rivest L-P. Statistical inference procedures for bivariate archimedian copulas. *Journal of the American Statistical Association* 1993; **88**: 1034–1043.

[13] Meester SG and MacKay J. A parametric model for clustered correlated categorical data. *Biometrics* 1994; **50**: 954–963.

[14] Frank MJ. On the simultaneous associativity of F(x,y) and x+y-F(x,y). *Aequationes Mathematicae* 1979; **19**: 194–226.

[15] Trégouet D-A, Ducimetière P, Bocquet V, Visvikis S, Soubrier F and Tiret L. A parametric copula model for analysis of familial binary data. *Am. J. Hum. Genet.* 1999; **64**: 886–893.

[16] Agresti A. Modeling ordered categorical data: recent advances and future challenges. *Statistics in Medicine* 1999; **18**: 2191–2207.

[17] Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**: 141–151.

[18] Joe H. Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis* 1993; **46**: 262–282.

[19] Mardia KV. *Families of Bivariate Distributions.* Griffin: London, 1970.

[20] Kimberling CH. A probabilistic interpretation of complete monotonicity. *Aequationes Math.* 1974; **10**: 152–164.

[21] Nelsen RB. *An Introduction to Copulas.*Springer-Verlag: New-York, 1999.

[22] Hoeffding W. Masstabinvariante korrelationstheorie. *Schriften des Matematischen Instituts und des Instituts fur Angewandte Methematik der Universitat Berlin* 1940; **5**: 179–233 [Reprinted in English as: Scale-invariant correlation theory in *The Collected Works of Wassily Hoeffding.* Fisher NI and Sen PK (eds), Springer Verlag: New-York, 57–107.

[23] Kruskal WH. Ordinal measures of association. *Journal of the American Statistical Association* 1958; **53**: 814–861.

[24] Kendall MG. The treatment of ties in ranking problems. *Biometrika* 1945; **33**: 239–251.

[25] Agresti A. The effect of category choice on some ordinal measures of association. *Journal of the American Statistical Association* 1976; **71**: 49–55.

[26] Darsow WF, Nguyen B and Olsen ET. Copulas and Markov processes. *Illinois Journal of Mathematics* 1992; **36**: 600–642.

Table 1: Selected one-parameter bivariate copula distributions.

Name $\theta \in$	$C_\theta(u,v)$ Limits and Special Cases
Frank $(-\infty,\infty)/\{0\}$	$-\frac{1}{\theta}ln[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}]$ $C_{-\infty}=W^{(a)}, C_0=\Pi^{(b)}, C_\infty=M^{(c)}$
Plackett $(0,\infty)/\{1\}$	$\frac{[1+(\theta-1)(u+v)]-\sqrt{[1+(\theta-1)(u+v)]^2-4uv\theta(\theta-1)}}{2(\theta-1)}$ $C_0=W, C_1=\Pi, C_\infty=M$
Clayton $[-1,\infty)/\{0\}$	$\max[(u^{-\theta}+v^{-\theta}-1)^{1/\theta},0]$ $C_{-1}=W, C_0=\Pi, C_\infty=M$
Joe $[1,\infty)$	$1-[(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta]^{1/\theta}$ $C_1=\Pi, C_\infty=M$
Mardia $[-1,1]$	$\frac{\theta^2(1+\theta)}{2}M(u,v)+(1-\theta)^2\Pi(u,v)+\frac{\theta^2(1-\theta)}{2}W(u,v)$ $C_{-1}=W, C_0=\Pi, C_1=M$

a:$W(u,v)=\max(u+v-1,0)$

b:$\Pi(u,v)=uv$

c:$M(u,v)=\min(u,v)$

Table 2: Incidence of observed pain severity scores per group.

Treatment	Pain Score	0.5h	1h	1.5h	2h	3h	4h	6h	24h
Placebo (N=12)	1 (no)	0	0	1	1	2	2	3	4
	2 (mild)	1	4	5	4	4	4	3	2
	3 (moderate)	6	5	2	3	2	3	3	0
	4 (severe)	5	3	4	4	2	0	0	0
5 mg (N=13)	1	1	0	1	1	4	5	6	9
	2	1	5	7	8	6	7	5	2
	3	10	8	5	4	3	0	0	0
	4	2	0	0	0	0	0	0	0
20 mg (N=14)	1	1	2	3	6	4	5	6	9
	2	3	7	8	6	10	6	5	3
	3	7	5	3	2	0	3	1	0
	4	3	0	0	0	0	0	0	0

Table 3: Kendall's tau-b coefficients for pain severity scores (no, mild, moderate, severe) measured over time across treatment groups (placebo, 5mg or 20mg).

	0.5h	1h	1.5h	2h	3h	4h	6h	24h
0.5h	1	**0.685**	0.538	0.560	0.312	0.155	0.186	-0.108
1h		1	**0.726**	0.704	0.519	0.276	0.370	0.02
1.5h			1	**0.808**	0.700	0.559	0.596	0.126
2h				1	**0.728**	0.582	0.637	0.230
3h					1	**0.768**	0.740	0.255
4h						1	**0.877**	0.458
6h							1	**0.553**
24h								1

Table 4: Log-likelihood for various dependence models and link functions. All cumulative marginal models have three thresholds along with group and log(time)as covariates.

Dependence	Logistic	Probit	Compl. Log-Log
Independent	-295.2	-293.5	-293.0
Frank	-190.0	-188.9	**-188.0**
Plackett	-191.8	-190.7	-189.6
Clayton	-197.9	-196.8	-192.3
Joe	-194.9	-194.4	-195.7
Mardia	-200.6	-199.5	-199.4
Exchangeable Frank	-226.1	-224.4	-224.1

Table 5: Parameter estimates from the stationary homogeneous first order model with Frank's copula and complementary log-log link.

Parameter	estimate (s.e.)
α_1	-3.14 (0.40)
α_2	-1.44 (0.30)
α_3	-0.01 (0.23)
group (5mg-pl)	1.19 (0.31)
group (20mg-pl)	1.10 (0.32)
Log(time)	0.81 (0.10)
θ	13.9 (2.15)

Table 6: Predicted Kendall's tau-b coefficients for each group over time.

0 mg	0.5h	1h	1.5h	2h	3h	4h	6h	24h
0.5h	1	0.72	0.57	0.45	0.36	0.29	0.23	*0.17*
1h		1	0.75	0.59	0.46	0.37	0.29	0.22
1.5h			1	0.76	0.58	0.46	0.36	0.27
2h				1	0.74	0.58	0.45	0.33
3h					1	0.75	0.58	0.42
4h						1	0.74	0.53
6h							1	0.68
24h								1
5 mg	0.5h	1h	1.5h	2h	3h	4h	6h	24h
0.5h	1	0.72	0.55	0.43	0.33	0.25	0.19	*0.08*
1h		1	0.74	0.57	0.43	0.33	0.25	0.11
1.5h			1	0.75	0.56	0.42	0.32	0.14
2h				1	0.73	0.55	0.41	0.18
3h					1	0.74	0.55	0.24
4h						1	0.73	0.32
6h							1	0.43
24h								1
20 mg	0.5h	1h	1.5h	2h	3h	4h	6h	24h
0.5h	1	0.72	0.55	0.43	0.33	0.25	0.19	*0.09*
1h		1	0.74	0.57	0.43	0.33	0.25	0.12
1.5h			1	0.75	0.56	0.43	0.32	0.15
2h				1	0.73	0.55	0.41	0.20
3h					1	0.74	0.54	0.26
4h						1	0.72	0.34
6h							1	0.46
24h								1

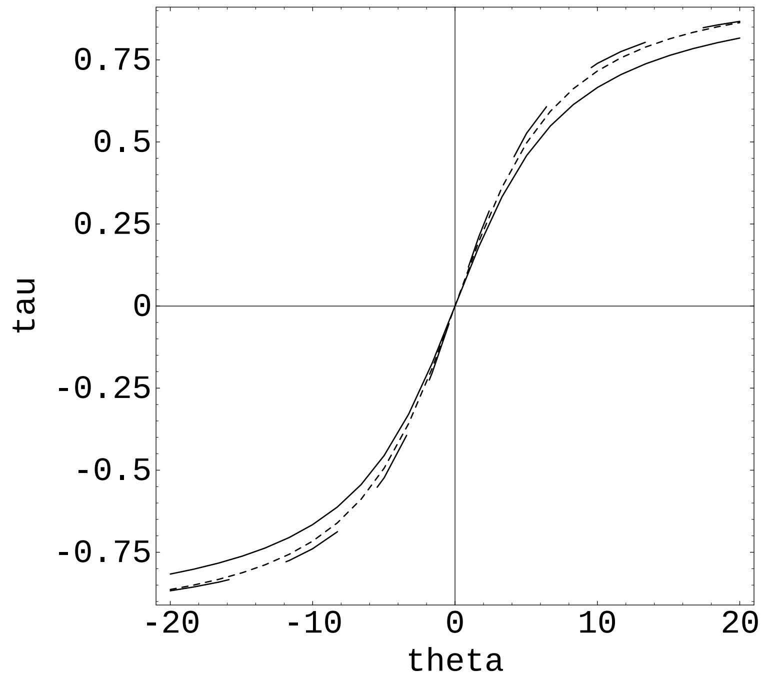

Figure 1: Relationship between Kendall's tau-b and the dependence parameter θ for the bivariate Frank's copula applied to two ordinal responses. Marginal uniform distributions ($\Pr[Y_j \leq k] = k/K$) are assumed, with K=4 (dashed) and with K=10 (dotted). Relationship for τ_C is also displayed (solid).

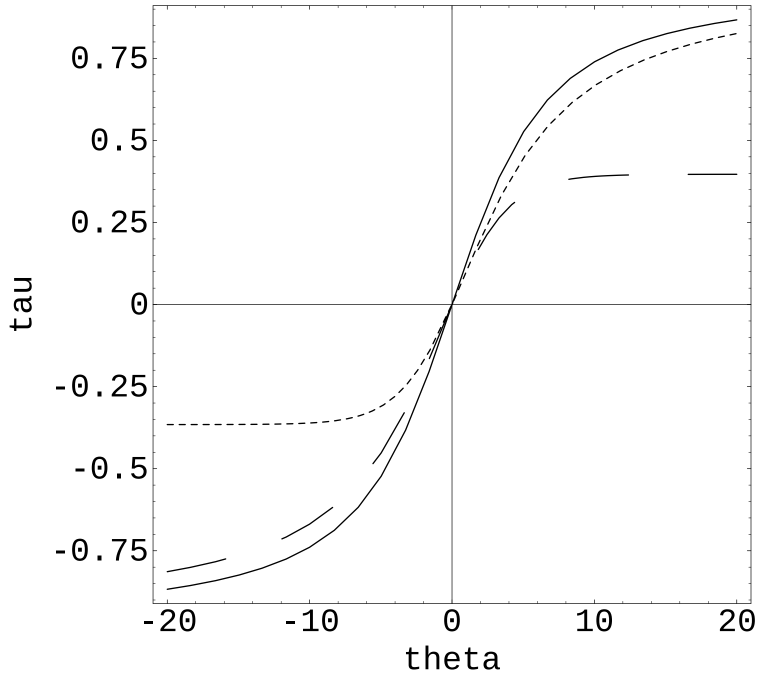

Figure 2: Relationship between Kendall's tau-b and the dependence parameter θ for the bivariate Frank's copula applied to two ordinal responses (K=4). Marginal distributions are $\Pr[Y_j \le k] = (k/K)^{\epsilon_j}$. Lines are displayed for uniform distributions: $\epsilon = (1, 1)$ (solid), for $\epsilon = (0.25, 0.25)$ (dotted) and for $\epsilon = (0.25, 4)$ (dashed).

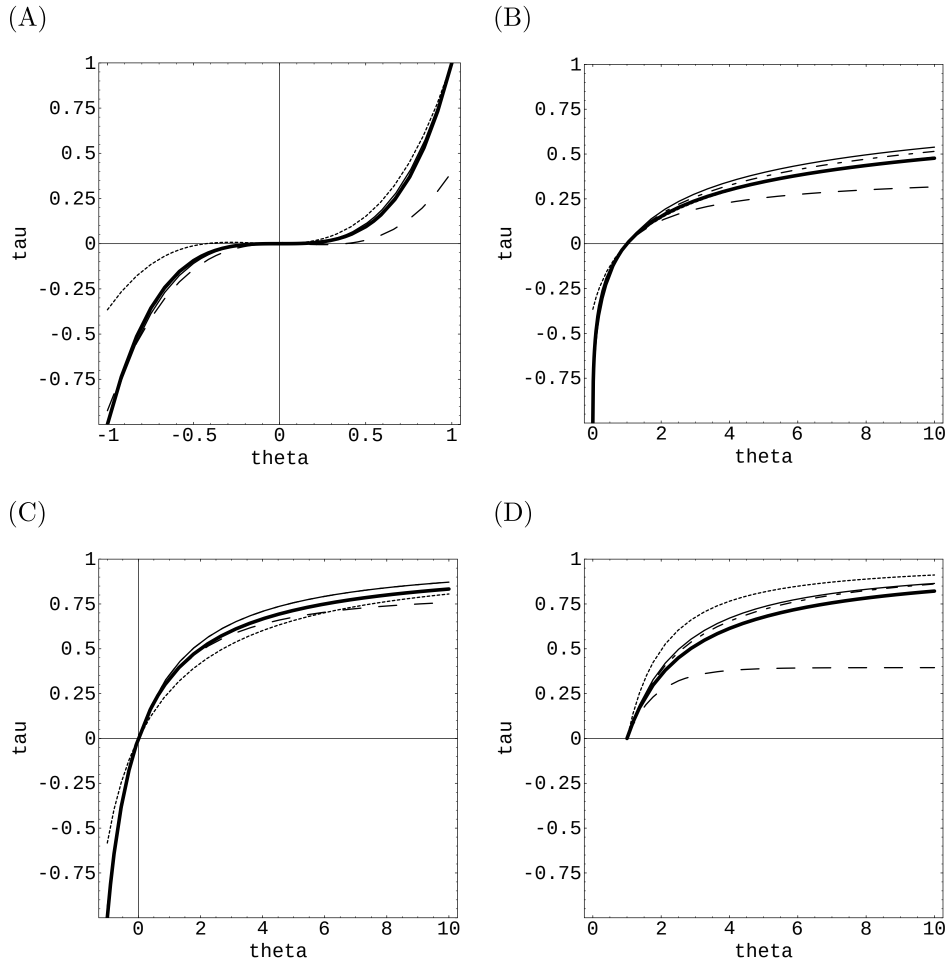

Figure 3: Relationship between the tau-b coefficient and the dependence parameter θ for bivariate copulas of Mardia (A), Plackett (B), Clayton (C) and Joe (D) applied to two ordinal responses. Marginal distributions are $\Pr[Y_j \leq k] = (k/K)^{\epsilon_j}$. Lines are displayed for uniform distributions: $\epsilon = (1, 1)$ with K=4 (solid) and with K=10 (-. -), for $\epsilon = (0.25, 0.25)$ with K=4 (dotted) and for $\epsilon = (0.25, 4)$ with K=4 (dashed). Relationship for τ_C is displayed in bold.

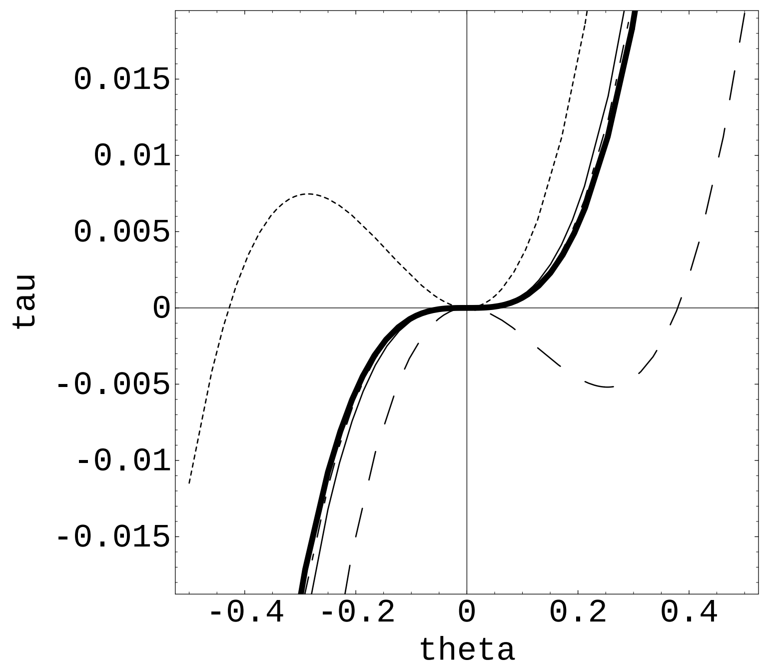

Figure 4: Subset of Figure 3A for Mardia's bivariate copula.