Quaderni di STATISTICA

VOLUME 13 - 2011

LIGUORI EDITORE

Volume 13, anno 2011

ISSN 1594-3739 (printed edition)

Registration n. 5264, 6 December 2001, Court of Justice at Naples
ISBN-13 **978 - 88 - 207 - 5619-2**

Director: Gennaro Piccolo

First italian edition, December 2011
Printed in Italy in December 2011 by OGL-Naples (IT)

This publication is financially supported by:
- Ministero delle Politiche Agricole, Alimentari e Forestali (gestione ex Centro per la formazione in economia e politica dello sviluppo rurale);
- Dipartimento di Economia, Università di Roma Tre
- Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, contributo PRIN 2007;
- Dipartimento TEOMESUS, Università di Napoli Federico II;

Quaderni di Statistica
Vol. 13, 2011

Modelling correlated bivariate ordinal data with CUB marginals

Marcella Corduas
Department TEOMESUS, University of Naples Federico II
E-mail: corduas@unina.it

Summary: The CUB model is a mixture distribution recently proposed in literature for modelling ordinal data. Although various methodological aspects of these models have been investigated, extensions are still needed in order to represent multivariate ordinal data. In this article we propose using the Plackett distribution in order to construct a one parameter bivariate distribution from CUB margins. The article examines both the methodological and computational issues and discuss the performance of the proposed technique by a simulation study.

Keywords: Ordinal data, CUB model, Multivariate distribution, Plackett's distribution.

1. Introduction

An alternative class of models, denoted CUB, has been recently introduced in order to model ordinal data originated by the assessment of a single or a group of items expressed by means of ratings or rankings (Piccolo, 2003; D'Elia and Piccolo, 2005).

The methodology has produced interesting results in several fields of applications ranging from linguistics (Balirano and Corduas, 2008) to medicine (D'Elia, 2008), from sociology (Iannario and Piccolo, 2010) to food marketing (Piccolo and D'Elia, 2008; Cicia et al. 2010; Manisera et al. 2011).

CUB model is a mixture distribution defined by the convex combination of a uniform and a shifted binomial distribution whose parameters may be related to explanatory variables characterizing raters or the object of evaluation. The model mimics a simplified choice mechanism which is supposed to underly the moulding of the judgements when a rater is requested to evaluate a certain item (such as a product or a service). The ratings are usually given by means of a Likert scale and concern rater's preferences, degree of satisfaction or, generally speaking, his/her agreement with a statement (Corduas *et al.*, 2009; Iannario and Piccolo, 2012).

Although various methodological aspects concerning CUB models have been investigated, extensions are still needed in order to represent multivariate ordinal data. For this aim we consider the method introduced by Plackett (1965) for constructing a one parameter bivariate distribution from given margins and propose using it for modelling ordinal data with CUB margins.

The article is organized as follows. Initially, the bivariate Plackett's distribution is illustrated. Then, assuming that the margins are described by a CUB distribution, the estimation problem is discussed. Finally, the performance of the proposed technique is assessed by a simulation study.

2. The Plackett distribution with CUB marginals

A bivariate Plackett random variable (X, Y) is characterized by a joint cumulative distribution function $H(x, y; \psi)$, $\psi \in (0, \infty)$, such that:

$$H(x, y; \psi) = \frac{M(x, y) - [M^2(x, y) - 4\psi(\psi - 1)F(x)G(y)]^{1/2}}{2(\psi - 1)}, \tag{1}$$

where $F(x)$ and $G(y)$ are the pre-defined marginal distribution functions defined on the support S_x and S_y, respectively (Plackett, 1965). Moreover, $M(x, y) = 1 + (F(x) + G(y))(\psi - 1)$ (Mardia, 1970). Statistical properties were investigated by Mosteller (1968), Steck (1968) and Wahrendorf (1980).

In literature various denominations involve Plackett distribution family. The reference can be either to the principle behind the genesis of the distribution or to specific features. For instance, this is the case of: "distribution with constant Yulean association", "contingency-type bivariate distribution" or "C-type distribution".

The parameter ψ is a measure of association between X and Y, in particular, $\psi = 1$ implies that X and Y are independent (so that $H(x, y; \psi) = F(x)G(y)$), whereas $\psi < 1$ and $\psi > 1$ refer to negative and positive association, respectively.

The distribution $H(x, y; \psi)$ satisfies the Fréchet bounds:

$$max\{F(x) + G(y) - 1, 0\} \leq H(x, y; \psi) \leq min\{F(x), G(y)\}, \tag{2}$$

where the lower and upper bounds are attained when $\psi \to 0$ and $\psi \to \infty$, respectively.

The original derivation of the Plackett distribution moves from considering the case of continuous margins and observing that one can always construct a joint cumulative distribution $H(x, y; \psi)$ such that when it is cut anywhere by lines parallel to the x and y axes, the probabilities in the four quadrants, viewed as a contingency table, lead to a cross-product ratio which remains constant for any choice of the cutting points (x, y).

This is a relevant constraint over the possible shapes that the joint distribution can assume. The problem goes back to the earlier contribution of Yule (1912), Pearson (1913), Pearson and Heron (1913) who lively debated about the probability model with constant association coefficient, its capability to reproduce the bivariate Normal and, therefore, to

model frequency surfaces in actual practice. The genesis of the distribution that Plackett introduced later, in 1965, is strictly related to that debate. The differences with the Normal distribution are mainly due to the fact that Plackett's model with Normal margins is characterized by the skewness of the conditional distributions, the nonlinearity of its regressions and, by definition, by the fact that the invariance property of the association coefficient is not verified (as follows by earlier Pearson's results, 1913; and Mosteller, 1968; Goodman, 1981).

Nevertheless, the Plackett distribution family has found numerous applications being the base for new types of models for continuous and discrete data. In the latter case, the distribution is taken as reference to the latent random variable from which a contingency table is derived by a discretization process. In this regards, overcoming the restriction on dimensions, which were initially limited to the bivariate or trivariate case, Molenberghs (1992) successfully extended the results to the multivariate Plackett distribution. Furthermore, Molenberghs and Lesaffre (1994) exploited that result for proposing a modelling approach to account the dependence of the association parameter from explanatory variables. Though marginal distributions are usually supposed to be continuous, the derivation of the Plackett distribution holds with convenient premises also in the discrete case.

In the rest of this article, we assume that (X, Y) is a discrete random variable with support $S_{xy} = \{(x, y) : x = 1, 2, ..., m; y = 1, ..., m\}$ and that the margins are described by CUB models. In particular, $X \sim F(x; \boldsymbol{\theta}_x)$ with $\boldsymbol{\theta}_x = (\pi_x, \xi_x)'$, and similarly $Y \sim G(y; \boldsymbol{\theta}_y)$, is characterized by the following distribution function:

$$F(x; \boldsymbol{\theta}_x) = \pi_x \sum_{j=1}^{x} \binom{m-1}{j-1} (1-\xi_x)^{j-1} \xi_x^{m-j} + (1-\pi_x)\frac{x}{m}, \qquad x = 1, 2, ..., m, \tag{3}$$

where $\xi_x \in [0, 1]$, $\pi_x \in (0, 1]$ and $m > 3$. The parameter space is therefore given by:

$$\Omega(\boldsymbol{\theta}_x) = \Omega(\pi_x, \xi_x) = \{(\pi_x, \xi_x) : \quad 0 < \pi \leq 1, 0 \leq \xi \leq 1\}. \tag{4}$$

The formulation of the corresponding probability mass distribution highlights the role of the two characterizing parameters. CUB model is, in fact, the mixture distribution:

$$p(x; \boldsymbol{\theta}_x) = \pi_x \binom{m-1}{x-1} (1-\xi_x)^{x-1} \xi_x^{m-x} + (1-\pi_x)\frac{1}{m}, \qquad x = 1, 2, ..., m. \tag{5}$$

The weight π_x determines the contribution of the Uniform distribution in the mixture, therefore, $(1 - \pi_x)$ is interpreted as a measure of the *uncertainty* which is intrinsic to any judgment. Besides, the parameter ξ_x, characterizes the shifted Binomial distribution and $(1 - \xi_x)$ denotes the degree of liking expressed by raters with respect to the item. In the former case $(1 - \xi_x) > 0.5$; the skewness is negative so that the portion of

raters which give a favourable judgement about the item under evaluation is large. The opposite is verified when $(1 - \xi_x) < 0.5$.

Computational issues were solved by Piccolo (2006) who provided an efficient algorithm for the maximum likelihood estimation of CUB models [1]. Further statistical properties and several extensions have been illustrated by Corduas (2011), Iannario and Piccolo (2010), and Iannario (2010; 2012) who proved that such a model is identifiable for $m > 3$ and improved its formulation in ordwer to account for *shelter* choices.

Given an observed sample of ordinal data, (y_i, x_i), for $i = 1, 2, ...n$, the estimation is performed by means of the two step procedure proposed by Joe and Xu (1996), the so called inference for the margins (IFM) method. Specifically, in the first stage only the parameters in the univariate margins, that is the CUB models are estimated by maximum likelihood. These step leads to: $\widehat{\boldsymbol{\theta}}_x$ and $\widehat{\boldsymbol{\theta}}_y$. The second stage involves maximum likelihood of the dependence parameter, ψ, with the univariate parameters held fixed from the first stage. The estimation, therefore, is performed by maximizing each of the following log-likelihood functions separately:

$$l_1(\boldsymbol{\theta}_x; \boldsymbol{x}) = \sum_{x=1}^{m} n_{x.} ln(p(x; \boldsymbol{\theta}_x)), \tag{6}$$

$$l_2(\boldsymbol{\theta}_y; \boldsymbol{y}) = \sum_{y=1}^{m} n_{.y} ln(p(y; \boldsymbol{\theta}_y)), \tag{7}$$

$$l_3(\psi; \boldsymbol{x}, \boldsymbol{y}) = \sum_{x=1}^{m} \sum_{y=1}^{m} n_{xy} ln(h(x, y; \psi)), \tag{8}$$

where, according to standard notation, n_{xy} is the frequency of the occurrence of (x, y) in the observed sample, $n_{x.}$ and $n_{.y}$ are the related marginal frequencies, and $h(x, y; \psi)$ is the probability mass distribution implied by (1).

In this respect, Joe (1997) showed that the IFM estimator is consistent, asymptotically Normal under regular conditions. In addition, Joe (2005) studied the asymptotic relative efficiency of IFM procedure compared with maximum likelihood estimation and considered some specific models indicating the typical level of efficiency.

2.1. The simulation experiment

A simulation study has been conducted in order to investigate the goodness of the Plackett distribution when fitting observations from a bivariate random variable having CUB margins.

[1] Iannario and Piccolo (2009) implemented the R code for the estimation of CUB models under various specifications. This is available at http://www.teomesus.unina.it/materiali/cub/

First, we briefly discuss a technique for simulating a bivariate discrete distribution with given margins and correlation. Specifically, we illustrate the approach proposed by Iman and Conover (1982) who introduced a distribution-free technique to generate observations from a set of random with a desired rank correlation matrix. Assuming that a sample of n observations from the (X, Y) random variable with rank correlation matrix $\boldsymbol{C}$ is needed, the procedure can be summarized as follows.

- A sample of n observations from an auxiliary random variable is generated. For instance, we generate a sample from a from $N(0, 1)$ random variable and collect the observations in a column vector $\boldsymbol{v} = (v_1, ..., v_n)'$;
- Construct the matrix $\boldsymbol{W} = (\boldsymbol{w_1}, \boldsymbol{w_2})$ where $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ are two column vectors obtained by random permutation of $\boldsymbol{v}$;
- Compute the transformation: $\boldsymbol{W}^*$=$(\boldsymbol{w_1^*}, \boldsymbol{w_2^*})$ so that $\boldsymbol{W}^*$=$\boldsymbol{WP}'$ where $\boldsymbol{C}$=$\boldsymbol{PP}'$;
- Generate a sample $\boldsymbol{x} = (x_1, ..., x_n)'$ from the marginal random variable X, and similarly generate $\boldsymbol{y} = (y_1, ..., y_n)'$ from Y;
- Rearrange the vector $\boldsymbol{x}$ so that it has the same ordering as the first column of the matrix $\boldsymbol{W}^*$, that is $\boldsymbol{w_1^*}$, and similarly rearrange $\boldsymbol{y}$ according to the ordering of $\boldsymbol{w_2^*}$. The resulting rearranged vectors will have the desired rank correlation.

The procedure is very simple to implement and, moreover, the choice of the auxiliary variable is rather flexible since it is not confined to any particular parametric distribution family.

Then, a simulation study has been performed considering the CUB models in Table 1. The parameter values have been selected so that various situations are represented. Examples A and B show two CUB distributions with rather similar pattern, whereas example C displays two distributions with opposite skewness (Figure 1). Note that, in order to facilitate the interpretation, the distributions in Figure 1 are illustrated by means of solid lines though they refer to discrete random variables.

Table 1. CUB parameters

Distribution	π_x	ξ_x	π_y	ξ_y
A	0.7	0.7	0.9	0.5
B	0.7	0.5	0.8	0.4
C	0.7	0.9	0.8	0.2

Moreover, the following values of rank correlation have been considered: $\rho_s = \{0.2, 0.5, 0.8\}$ for the generation of samples from the bivariate population (X, Y) with given CUB margins. It is reasonable to expect that the results will be symmetric with respect to the null correlation; for this reason we have selected only positive values of

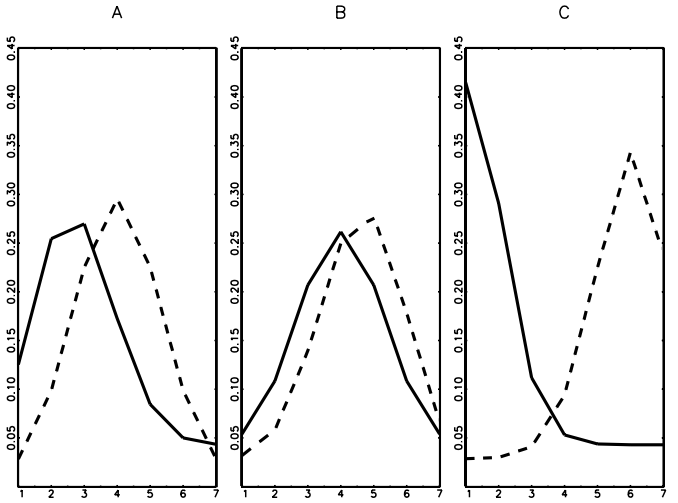

Figure 1. Marginal probability mass distribution of X (=solid line) and Y (=dashed line)

ρ_s. Then, the experiment has been conducted according to the following plan where the simulation was iterated 500 times.

Having selected the values of the CUB parameters and of the correlation ρ_s:

- two samples $\boldsymbol{x}$ and $\boldsymbol{y}$, with n=1000, are independently generated by the two CUB models representing the marginal distributions;
- Iman and Conover procedure is applied in order to rearrange the two samples so that the requested rank correlation is obtained;
- a CUB model is estimated for each sample leading to $F(x;\widehat{\boldsymbol{\theta}}_x)$ and $G(y;\widehat{\boldsymbol{\theta}}_y)$, respectively ;
- the parameter ψ is estimated by maximum likelihood according to (8).

The goodness of fit of the estimated bivariate Plackett distribution is evaluated by means of the following normalized dissimilarity index:

$$\Delta = \frac{1}{2}\sum_{x=1}^{m}\sum_{y=1}^{m}\left|h(x,y;\widehat{\psi}) - \frac{n_{xy}}{n}\right|. \tag{9}$$

It measures the amount of probability mass that has to be moved from one cell to the others so that the perfect match of the estimated joint distribution with the empirical one is achieved.

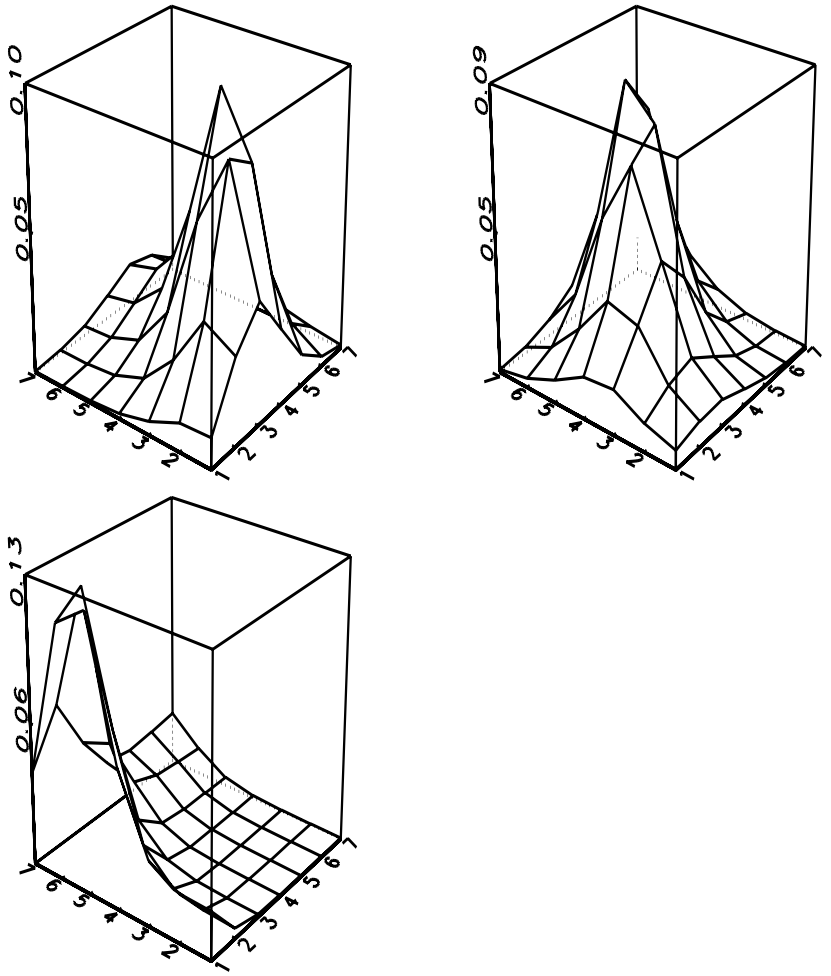

Figure 2. Joint probability distributions with CUB marginals (by rows: A,B,C)

In order to give an insight into the shape of the joint probability distributions implied by the three examples, in Figure 2 we illustrate the theoretical Plackett distributions with ψ held fixed to the average value obtained from the simulation study. The ψ value is in fact strictly dependent from the selected rank correlation and the average value from simulation is rather stable across the three cases.

The normalized dissimilarity index is on average rather small and ranges between 0.06 and 0.08 (Table 2). In addition, for the case B with $\rho_s = 0.5$ we considered the contribution to the dissimilarity index originated by each cell of the 7x7 table which is

Table 2. Dissimilarity index (average)

Distributions \ ρ_s	0.2	0.5	0.8
A	0.0737	0.0793	0.0791
B	0.0764	0.0831	0.0879
C	0.0624	0.0669	0.0728

generated at a given iteration of the simulation study. Specifically, we evaluated:

$$0.5 \left| h(x, y|\widehat{\psi}) - \frac{n_{xy}}{n} \right|$$

for each couple (x, y).

Figure 3 illustrates the box-plots obtained for the 500 values generated in correspondence of one of those cells. The box-plots are collected by columns so that the first panel refers to the cells in the first column of the 7x7 table and so on.

In particular, the box-plots highlight that the empirical and fitted distribution are generally close. The worst fitting is obtained about the peak of the distribution where the dissimilarity shows larger variability. However, the median ranges between 0.00020 and 0.0033, and the third quartile goes from 0.00018 to 0.0053, proving the effectiveness of the overall performance of the proposed technique.

3. Final remarks

This article is the first study aimed at finding a framework to extend CUB models to the multivariate case. The results achieved are encouraging.

Despite the known limits, the Plackett distribution seems to fit well the joint distribution having CUB margins.

However, some further investigation is needed. Firstly, we need to solve the computational problems involved with multiple random variables. Secondly, the model has to be extended in order to include covariates: this is in fact one of the point of strength of CUB models. Finally, more flexible probability models should be considered in order to remove some of the distributional constrains implied by the constancy of the association parameter which is fundamental for the Plackett distribution.

Acknowledgements: This work has been supported by MUR PRIN2008 grant CUP E61J10000020001 – University of Naples Federico II.

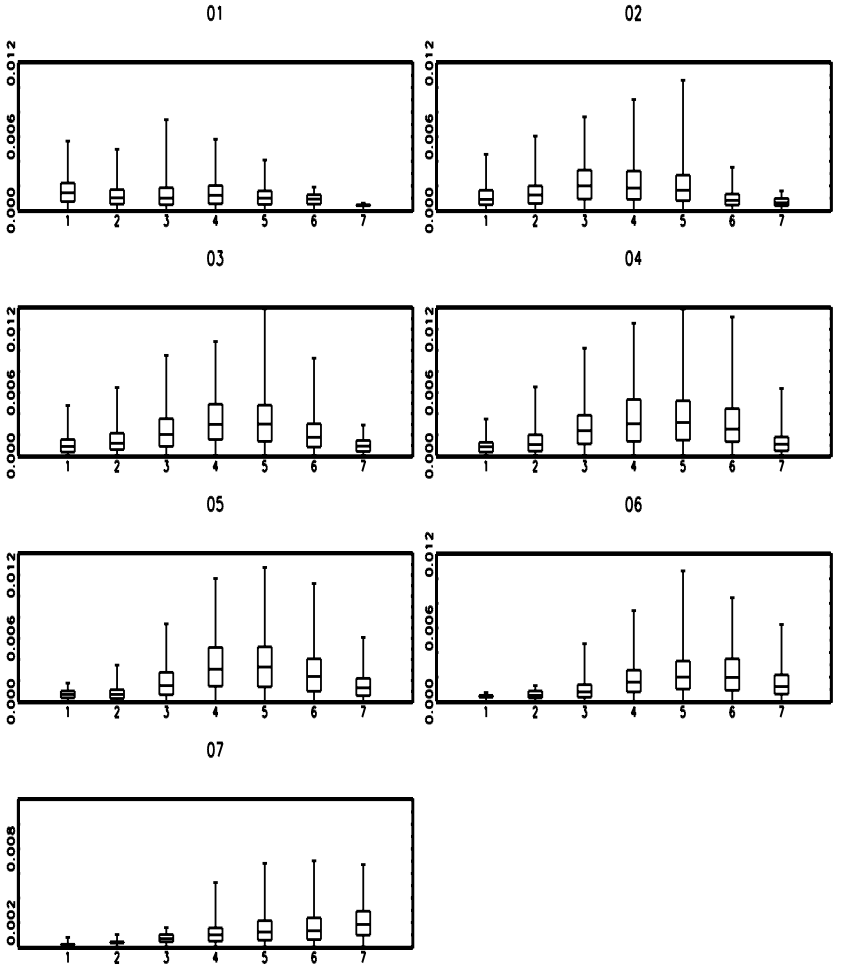

Figure 3. Box plots of the absolute differences between empirical and fitted distribution for each cell of the 7x7 contingency table

References

Balirano G., Corduas M. (2008), Detecting semiotically expressed humor in diasporic TV productions, *HUMOR: International Journal of Humor Research*, 3, 227-251.

Cicia G., Corduas M., Del Giudice T., Piccolo D. (2010), Valuing consumer preferences with the CUB model: a case study of fair trade coffee, *International Journal On Food System Dynamics*, 1, 82–93.

Corduas M. (2011), Assessing similarity of rating distributions by Kullback-Liebler divergence, in B. Fichet, D. Piccolo, R. Verde, M. Vichi (Eds), *Classification and Multivariate Analysis for Complex Data Structures*, Springer-Verlag, Heidelberg, 221–228.

Corduas M., Iannario M., Piccolo D. (2009), A class of statistical models for evaluating services and performances, in M. Bini, P. Monari, L. Salmaso (Eds), *Statistical methods for the evaluation of educational services and quality of products*, Physica-Verlag, Heidelberg, 99–117.

D'Elia A. (2008), A statistical modelling approach for the analysis of TMD chronic pain data, *Statistical Methods in Medical Research*, 17, 1–15.

D'Elia A., Piccolo D., (2005), A mixture model for preference data analysis, *Computational Statistics and Data Analysis*, 49, 917–934.

Goodman L.A. (1981), Models and the bivariate normal for contingency tables with ordered categories, *Biometrika*, 68, 347–355.

Iman, R. L., Conover, W. J. (1982), A distribution- free approach to inducing rank correlation among input variables, *Communications in Statistics*, Bl1, 311–334.

Iannario M. (2010), On the identifiability of a mixture model for ordinal data, *METRON*, LXVIII, 87–94.

Iannario M. (2012), Modelling shelter choices in a class of mixture models for ordinal responses, *Statistical Methods and Applications*, 21, 1–22.

Iannario M., Piccolo D. (2009), A program in R for CUB models inference, *http://www.teomesus.unina.it /materiali /cub*

Iannario M., Piccolo D. (2010), Statistical modelling of subjective survival probabilities, *GENUS*, 66, 17–42.

Iannario M., Piccolo D. (2012), CUB models: Statistical methods and empirical evidence, in: Kenett R. S. and Salini S. (eds.), *Modern analysis of customer surveys: with applications using R*, J. Wiley & Sons, Chichester, 231–258.

Joe H. (1997), *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

Joe H. (2005), Asymptotic efficiency of the two-stage estimation method for copula-based models, *Journal of Multivariate Analysis*, 94, 401–419

Joe, H. and Xu, J.J. (1996), The estimation method of inference functions for margins for multivariate models, *Technical Report n. 166*, Department of Statistics, University of British Columbia.

Manisera M., Piccolo D., Zuccolotto P. (2011), Analyzing and modelling rating data for sensory data in food industry, *Quaderni di Statistica*, 13, 69–82.

Mardia K.V. (1970), *Families of bivariate distributions*, Griffin, London.

Molenberghs G. (1992), *A multivariate Plackett distribution with given marginal distributions*, Universitaire Instelling Antwerpen, n.92/33.

Molenberghs G., Lesaffre E. (1994), Marginal modelling of correlated ordinal data using multivariate Plackett distribution, *Journal of the American Statistical Association*, 89, 633–644.

Mosteller F. (1968), Association and estimation in contingency tables, *Journal of the American Statistical Association*, 63, 1–28.

Pearson K. (1913), On the theory of association, *Biometrika*, 9, 159–315.

Pearson K., Heron D. (1913), Note on the surface of constant association, *Biometrika*, 9, 534–537.

Piccolo D. (2003), On the moments of a mixture of Uniform and shifted Binomial random variables, *Quaderni di Statistica*, 5, 85-104.

Piccolo D. (2006), Observed information matrix for MUB models, *Quaderni di Statistica*, 8, 33–78.

Piccolo D., D'Elia, A. (2008), A new approach for modelling consumers' preferences, *Food Quality and Preferences*, 19, 247–259.

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

Steck G.P. (1967), A note on contingency-type bivariate distributions, *Biometrika*, 55, 262–264.

Wahrendorf J. (1980), Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions, *Biometrika*, 67, 15–21.

Yule G. U. (1912), On the methods of measuring association between two attributes, *Journal of the Royal Statistical Society*, 75, 579–642.