A proposal for the multidimensional extension of CUB models

Federico Andreis and Pier Alda Ferrari

Abstract Particular emphasis has been put, lately, on the analysis of categorical data and many proposals have appeared, ranging from pure methodological contributions to more applicative ones. Among such proposals, the CUB class of distributions, a mixture model for the analysis of ordinal data that has been successfully employed in various fields, seems of particular interest. CUB are univariate models that do not possess, at present, a multivariate version: aim of the present work is to investigate the feasibility of building a higher-dimensional version of such models and its possible applications. In order to achieve such results, we propose to employ techniques typical of the framework of copula models, that have proven to be valid tools for multivariate models construction and data analysis.

Key words: multivariate ordinal data, CUB models, copula models, dependence structures

1 Introduction

The analysis of ordinal data is nowadays a field of great interest for the vast majority of applied fields and poses interesting challenges to statisticians in the development of an adequate methodology. Diverse proposals have been introduced during the recent years for their treatment, leading to important theoretical contributions from the scholars worldwide. Among such proposals, the authors deem worth of particular consideration the CUB [5] models, a class of univariate mixture distributions that has been successfully applied in many fields such as semiotics, ability assessment,

Federico Andreis
Università degli Studi di Milano, e-mail: federico.andreis@unimi.it

Pier Alda Ferrari
Università degli Studi di Milano, e-mail: pieralda.ferrari@unimi.it

medical research and customer satisfaction; the parsimonious parameterization and the ease of estimation and interpretation make CUB models a very useful tool for ordinal data analyses. Our proposal aims at extending such modeling approach to be able to handle multivariate data, the main reason for it being the belief that multivariate data coming from the same source (such as responses to a questionnaire items) should be treated as an ensemble, rather than split up, in order to account for (possibly) existing dependence structures in the estimation procedure, analysis and final results interpretation. Such extension is sought for in the framework of copulas [4], with the awareness of the problems arising from working with categorical, rather than continuous, data; a general structure is proposed and the use of different copula models is investigated.

2 Background

This section is intended to briefly review the general framework of both CUB and copula models.

2.1 CUB models

The CUB is a class of mixture models, possibly involving covariates, developed as a new approach for modeling discrete choices processes. The most common situation in which such approach can be employed regards the analysis of questionnaire data, with items responses evaluated on Likert scales and, thus, in the presence of ordinal data. The inherent *uncertainty* component is modeled through a discrete uniform variable, whereas the latent process leading to the choice, and governed by the subjective *feeling*, is modeled using a Shifted-Binomial distribution. The probability of observing a particular response r to an item, assuming that the number of item categories m is known and fixed, is expressed as a mixture of two such components as follows:

$$P(R=r) = \pi \binom{m-1}{r-1} \xi^{m-r}(1-\xi)^{r-1} + (1-\pi)\frac{1}{m}, \quad r = 1,2,...,m \tag{1}$$

with $\pi \in (0,1]$ and $\xi \in [0,1]$.
π define the mixture weights and as such is inversely related to the amount of *uncertainty* in the answers (the higher π, the less the uniform component contributes to the mixture), i.e. *"each respondent acts with a **propensity** to adhere to a thoughtful and to a completely uncertain choice, measured by π and $(1-\pi)$, respectively"* [3].
ξ, on the other hand, is related to personal preferences and measures the strength of *feeling* or *adherence*, *agreement* with the item (the interpretation of ξ also depends on the kind of ordering adopted for the item).

2.2 Copula models

Copulas are n-place, grounded and n-increasing real functions with the unit hypercube as domain, that can be used to link univariate distribution functions (called margins) to form multivariate distribution functions according to arbitrary dependence structure (for reference see, for example, [4]). One of the main advantages of copulas is that they allow for separate specification of margins and dependence among them, where the dependence structure can be (and usually is) characterized by one or more parameters. Sklar's Theorem [6] is central to the theory of copulas: it provides the representation through a copula of a multivariate distribution function and grants its uniqueness when dealing with continuous margins. Such fact grants many useful properties that can be exploited for estimation and inferential purposes.

Particular care is needed when working with discrete margins, as in the case of the CUB models, due to the non-uniqueness issue, as stressed in [2]; non-uniqueness stems from the fact that marginal distribution functions are not strictly monotonically increasing, rather monotonically non-decreasing, and do not possess an inverse in the usual sense, rather a pseudo-inverse (see, for example, [4]). The most severe consequence of this is that it becomes impossible to draw general conclusions on the dependence structure binding the margins based on the copula parameterization alone (every result in this sense has been shown to be margin-dependent). Nonetheless, copulas still are an easy-to-implement and interesting tool to build multivariate models, and under certain circumstances it is still possible to make assessments about dependence among margins. For example, some copula families possess the property of being ordered by Positive Quadrant Dependence (PQD, see [2]): this grants minimal requirements to be met for copula parameters to be interpretable as dependence measures. We will therefore focus on such families in this work.

3 Proposal

As said, CUB models have been developed to describe univariate discrete phenomena, e. g. the distribution of answers to a single questionnare item. Since questionnaires are usually composed by many different questions (say k), a complete analysis with CUB would require to separately estimate the k couples $(\pi_i, \xi_i), i = 1, ..., k$, that characterize each item. This disjoint analysis approach does not take into account the dependence (possibly) existing among items, which could be exploited to better catch further information about the phenomenon and enrich its understanding. Drawing on this, we intend to evaluate the feasibility of a multivariate joint approach to CUB modeling, through the use of copula models. We thus define a multidimensional extension of CUB models, and call it CO-CUB model, as follows:

Definition 1. A k-dimensional ($k \geq 2$) CO-CUB model with copula C is a multivariate discrete variable with margins $R_i \sim CUB(\pi_i, \xi_i)$, $i = 1, ..., k$, each with support $\{1, ..., m_i\}$, $m_i > 3$, and joint distribution function given by:

$$\begin{aligned}\Psi(r_1,...,r;\underline{\pi},\underline{\xi},\underline{\theta}) = P(R_1 \leq r_1,...,R_k \leq r_k;\underline{\pi},\underline{\xi},\underline{\theta}) = \\ = C_{\underline{\theta}}[F_1(r_1;\pi_1,\xi_1),...,F_k(r_k;\pi_k,\xi_k)]\end{aligned} \qquad (2)$$

where $\underline{\pi} = (\pi_1,...,\pi_k)'$, $\underline{\xi} = (\xi_1,...,\xi_k)'$ and for a particular choice of copula C, characterized by a parameter $\underline{\theta} = (\theta_1,...,\theta_d)'$ taking values in some real d-dimensional space Θ defining the dependence structure of its components. $F_i(r_i) = F_i(r_i;\pi_i,\xi_i)$ stands for the distribution function of the i-th margin, i.e. $F_i(r_i) = P(R_i \leq r_i)$, and the support of the CO-CUB variable is the grid $\{1,...,m_1\} \times ... \times \{1,...,m_k\}$. The whole parameter set for a k-dimensional CO-CUB is, then, the ordered triplet $(\underline{\pi},\underline{\xi},\underline{\theta}) \in (0,1]^k \times [0,1]^k \times \Theta$.

An interesting first attempt at defining a bivariate CUB distribution using the Plackett distribution is made in [1]. Our proposal further extends this approach to a more general framework, focusing on the comparison of different choices of the copula C to define the CO-CUB models, of which [1] is shown to be a special case; while developing a general method, we specifically compare, for the sake of illustration, the well known Clayton, Frank and Plackett copulas (whose families are ordered by PQD) in the simple bivariate case, discussing from a methodological point of view estimation-related issues and parameters interpretation, as well as feasibility of extension to more than $k = 2$ dimensions.
By definition of copula, margins of (2) are all CUBs, whose parameters retain, then, the same interpretation as in the unidimensional case, while for what concerns the copula parameter θ, its interpretation as a dependence measure will be a further subject of studying: a first idea is to use it to rank couples of items by strength of dependence, when in presence of an ordered family of copulas (previously described). This might be useful for questionnaire calibration purposes and to individuate latent structures.

Acknowledgements The authors acknowledge financial support from the European Social Fund Grant (FSE), Lombardy Region.

References

1. Corduas, M.: Modelling correlated bivariate ordinal data with CUB marginals. Quaderni di Statistica **13**, 109–119 (2011)
2. Genest, C. and Neslehova, J.: A Primer on Copulas for Count Data. ASTIN Bulletin vol.37 no.2 (2007). http://www.actuaries.org/LIBRARY/ASTIN/vol37no2/475.pdf
3. Iannario, M. and Piccolo, D.: A program in R for CUB models inference (Version 2.0), (2009) available at: http://www.dipstat.unina.it
4. Nelsen, R.B.: An Introduction to Copulas. Springer (2010)
5. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. Quaderni di Statistica **5**, 85–104 (2003)
6. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris, **8**, 229–231 (1959)