JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

A NEW PARADIGM FOR MODELLING ORDINAL RESPONSES IN SAMPLE SURVEYS[1]

Maria IANNARIO[2]
Department of Political Sciences, University of Naples Federico II

E-mail: maria.iannario@unina.it

Domenico PICCOLO[3]
Department of Political Sciences, University of Naples Federico II

E-mail: domenico.piccolo@unina.it

Abstract:

A growing interest in the current surveys is focused on human and relational issues collected as ordinal variables. Standard approaches interpret them as manifest expressions of a continuous latent variable and the current methodology is based on the relationship between the cumulative function of the ratings and the subjects' variables. A different class of models, called CUB, is based on the statement that ordinal responses are a weighted combination of a personal feeling and an inherent uncertainty surrounding the decisional process. In this paper, the novel paradigm is presented and applied to real data sets to show the advantages of this method for analyzing big data in the context of official statistics.

Key words: Ordinal responses, Cumulative models, CUB models

1. Introduction

The paper is concerned with the analysis of ordinal variables which are common in official statistical surveys. Some relational variables such as happiness, job satisfaction, quality of life, trust in the others, etc. are frequently considered as the main responses. They are collected as variables expressed on a discrete ordinal scale and are characterized by phenomena where several factors affect human behaviour, in connection or apart from the usual economic variables [3, 20]. Their study is justified by the awareness that the human well-being is an essential component of the economic development and it represents an important indicator of economic performance and social progress [21].

Current methodologies include the study of these data in the context of Generalized Linear Models [17]. They assume that the discrete response is obtained by grouping the latent variable surrounding individual choice in classes of values by means of cutpoints. Moreover, they are based on the relationship between the cumulative function of the ratings and the subjects' variables. However, the departure from this usual practice could be necessary. First, because it is often difficult to summarize and visualize hundredths of expressed scores on several items by using plots and functions which are not immediately related to the latent constructs and second, because the estimation of several cutpoints worsens the model parsimony.

This paper tackles a different approach [18]. It is based on the direct analysis of the mechanism of choice with some advantages in the estimation process [19] since it adheres to latent variables paradigm without the need to estimate cutpoints. This line of reasoning may be convenient when ordinal responses are collected and visualization and communication are important objectives.

The framework denoted as CUB models is useful for a parametric assessment of the psychological process of selection of a grade on a Likert scale. It weights the two main latent components that characterize selection: the feeling expressed by an individual and the uncertainty which marks out the selection.

The recent interest in well-being and happiness measurements has inspired the application of this class of models for the selection of response categories in a number of several research areas related to these topics ([4, 5], for instance). There are also examples in other contexts. A comprehensive reference list is presented in [14] where it is possible also to refer for the estimation of the models by means of the open source R environment [16]. Notation and inferential aspects have been carried out in [12].

In the next section, notations and extensions are proposed. Then, in section 3 the description of case studies is presented in order to show the advantages of this method for analysing big data in the context of official statistics. Finally, some conclusions end the paper.

2. Specification and extension of CUB models

The mixture model we will propose is motivated by the circumstance that people transform own internal perception into an expressed score according to a given ordinal sequence of categories. This mixture may consist of a Combination of discrete Uniform and shifted Binomial random variables (CUB). It mimics the uncertainty in the process selection and the motivations derived by individual characteristics/background. Main aspects concerning the link among the two main components and the chosen probability distributions have been proposed in [10].

Briefly, the Uniform distribution is considered because it is the most extreme and uninformative case among the discrete random variables. Instead, the shifted Binomial is used as an approximation of a counting process of selection among categories, in the sense that each response may be interpreted as the result of cumulated choices against different alternatives.

Formally, given explanatory variables t ∈ T , let Yi be the ordinal response take values in {1, 2, . . . , m}. Then, the CUB mixture has been defined for each respondent by:

$$Pr(Y_i = j \mid \mathrm{C}_i, \theta) = \pi_i\, b_j(\xi_i) + (1-\pi_i)\, p_j{}^U,\ j = 1, 2, \ldots, m. \qquad (1)$$

We set $\mathrm{C}_i = (y_i, \mathrm{t}_i)$ the information set; $b_j(\xi_i) = \binom{m-1}{j-1} \xi_i{}^{m-j} (1-\xi_i)^{j-1}$ and $p_j{}^U = 1/m$, for $j = 1, 2, \ldots, m$, the probability mass functions of the shifted Binomial and discrete Uniform random variables, respectively. If we consider the information on subjects' covariates extracted from T and a logistic link used to preserve the mapping between parameters and covariates, we have:

$$logit(\pi_i) = \mathrm{x}_i\beta\ ; \qquad logit(\xi_i) = \mathrm{w}_i\gamma; \qquad i = 1,2,...,n.$$

Here, $\mathrm{t}_i, = (\mathrm{x}_i{}', \mathrm{w}_i{}')'$ is the information set useful to specify the relationship of π_i and ξ_i with the corresponding subjects' covariates x_i and w_i. Given the chosen parameteri-

zation, the covariates in x_i and w_i may coincide, overlap or be distinct. Then, the parameter vector $\theta = (\beta', \gamma')'$ is split with respect to the impact of uncertainty and feeling components, respectively.

In case of multi-items, the structure may be extended with the inclusion of objects' or contexts' characteristics of the $h = 1, 2, \ldots, H$ items. The K covariates $z_h = (z_{h1}, z_{h2}, \ldots, z_{hK})$ related to the h-th context imply that each row vector of the model is replicated n_h times for $i = 1, 2, \ldots, nh$; and $h = 1, 2, \ldots H$,

$$(y_i^{(h)} | 1, x_{i1}^{(h)}, x_{i2}^{(h)}, \ldots, x_{ip}^{(h)} \quad | 1, w_{i1}^{(h)}, w_{i2}^{(h)}, \ldots, w_{iq}^{(h)} | z_{h1}, z_{h2}, \ldots, z_{tK}).$$

Thus, we have:

$$logit(\pi_i^{(h)}) = x_i^{(h)} \beta + z_h \nu; \qquad logit(\xi_i^{(h)}) = w_i^{(h)} \gamma + z_h \eta;$$

where $\nu = (\nu_1, \nu_2, \ldots, \nu_K)'$ and $\eta = (\eta_1, \eta_2, \ldots, \eta_K)'$ are the parameter vectors which measure the impact of the context characteristics on *uncertainty* and *feeling* components, respectively. For analysing the possibility of random effects caused by the group membership on individual behaviour, a hierarchical CUB model (HCUB) has been proposed [9].

To interpret the standard CUB model (1) we consider the probability distribution for a given subject by letting $\pi_i = \pi$ and $\xi_i = \xi$. It can be considered as a global model which gives a synthetic measure of feeling and uncertainty for the whole sample of respondents.

$$Pr(Y = j \mid \theta) = \pi b_j(\xi) + (1 - \pi) p_j^U, \; j = 1, 2, \ldots, m, \qquad (2)$$

where $\theta = (\pi, \xi)'$ and the parameter space is the unit square. The identifiability of the model has been proved [7] for any $m > 3$; notice that $m = 3$ implies a saturated model.

According to (2), each respondent acts with a *propensity* to adhere to a thoughtful and to a completely uncertain choice with weights measured by (π) and ($1 - \pi$), respectively.

Thus ($1 - \pi$) is a measure of uncertainty. The level of *feeling*, instead, a component which needs to be specified on the basis of the survey, may be interpreted as a measure of agreement towards the item and it is measured by ($1 - \xi$). Then, the visualization of the models are shown as points in correspondence with ($1 - \pi$, $1 - \xi$).

Moreover, if a subset of respondents selects a specific option to simplify a more demanding choice it is possible to consider the extension of CUB models with a *shelter effect* [8]:

$$Pr(Y = j \mid \theta) = \delta[D_j^{(c)}] + (1 - \delta)[b_j(\xi) + (1 - \pi) p_j^U] \qquad j = 1, 2, \ldots, m. \quad (3)$$

For a given c, the presence of a possible *shelter* effect is introduced by a dummy variable $D_i^{(c)}$ which is 1 if $j = c$ and 0 otherwise. This circumstance is quite frequent when respondents are attracted by a peculiar wording of the questionnaire or when they would avoid critical options, for instance. The extension with inclusion of covariates for all parameters in (3) has been also implemented leading to GeCUB models [15].

Inference of CUB models is obtained by means of Maximum likelihood (ML) theory [19]. Specifically, ML estimates are obtained by the EM algorithm, whereas fitting measures are based on deviance and BIC criterion, among others. A dissimilarity index, which represents the proportion (=relative frequency) of subjects to move among the cells of the frequency distribution to achieve a perfect fit, is a very useful measure.

Extensions and generalizations of this class of models concern also variants of univariate distributions and of the probability distributions of components. In this respect we mention CUBE models [10, 11] which allow to capture overdispersion and CUB models with varying uncertainty [6].

A multivariate approach for the joint modelling of items has been pursued by means of a multi-objects approach [20] and via copula functions [1]. Multivariate CUB models via latent variables approach is an alternative task under scrutiny.

3. Empirical analysis

For presenting the main characteristics of the approach we consider two data sets which stem from big surveys. In the first case the standard model with significant covariates is involved to show the usefulness of the approach for the visualization of the results. In the second case an extension to the contextual structure, commonly present in official statistics, is found to be significant.

3.1 *Perceived happines* from SHIW data

Data stem from the Survey on Household Income and Wealth (SHIW) freely available on http://www.bancaditalia.it/statistiche/indcamp.

The survey conducted since 1965 by the Bank of Italy collects several information on the economic behaviour of Italian households. Specifically it measures income and wealth components. It also gathers information regarding job, health, perceived happiness, economic perceived conditions, family choices, capital gains, inheritance, financial information, and so on. Details on the survey design and on the content of the questionnaire can be found in [2]. In this context, the *perceived happiness* is expressed on a Likert scale from 1 to 10 by means of the analysis of respondents' behaviour and characteristics. We refer to 2012 wave with a validated sample of 8, 148 respondents.

In Figure 1, the observed distributions of relative frequencies and the estimated probabilities by CUB models are shown. Standard and CUB models with *shelter effect* are represented in left and right panels, respectively. From the empirical distribution a concentration of score at category 8 can be detected; thus, a sensible improvement is obtained by fitting the model with a *shelter choice* at $c = 8$. The dissimilarity index, which compares observed and fitted distributions, decreases from 0.075 to 0.048.

Fig. 1 CUB model (left) and CUB model with *shelter effect* (right) for *perceived happiness*

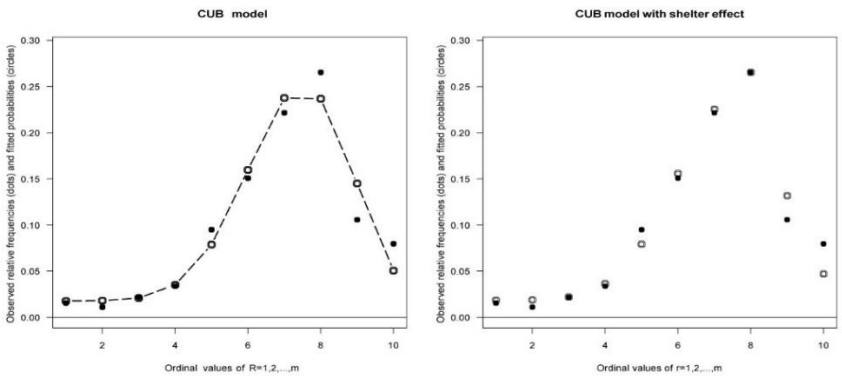


Table 1 summarizes the estimation of both models. It points out a high level of *perceived happiness* with a low uncertainty in this survey and improved fitting results for the second model.

Table 1. Estimation of CUB model (left) and CUB model with shelter effect (right) for *perceived happiness*.

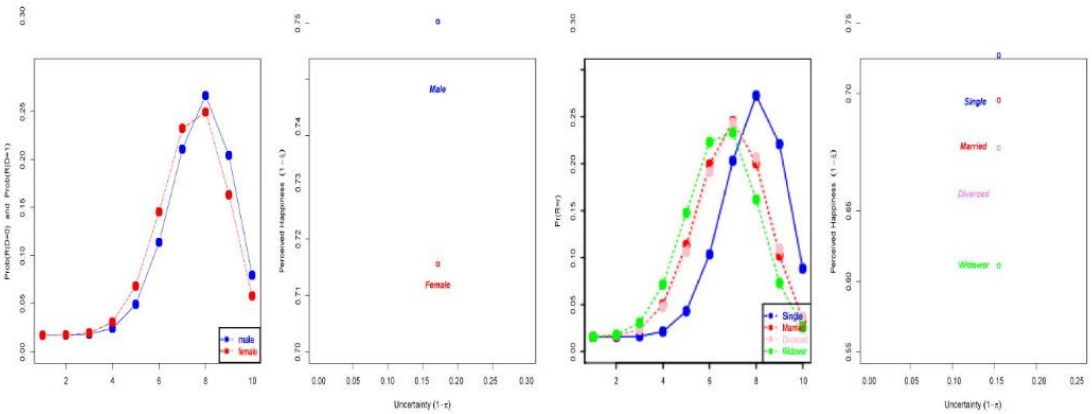
Models	Uncertainty parameters	Feeling parameters	*Shelter* parameter	$\ell(\theta)$	*BIC*
CUB	$\hat{\pi} =$ 0.823 *(0.008)*	$\xi =$ 0.301 *(0.002)*		−16005.95	32029.90
CUB+*shelter*	$\hat{\pi} =$ 0.807 *(0.008)*	$\xi =$ 0.307 *(0.002)*	$\delta =$ 0.046 *(0.007)*	−15983.40	31993.83

For a better understanding of this class of models, the introduction of covariates is considered. First, we introduce a *gender* variable for the feeling component with a constant uncertainty. On the first panel of Figure 2 we report the estimated distributions of men and women (men are happier than women) which are represented (second panel) by two points in the parameter space (higher position implies higher feeling). These results are summarized by:

$$logit(\pi) = 0.828; \qquad logit(\xi_i) = -1.100 + 0.177\, gender_i.$$

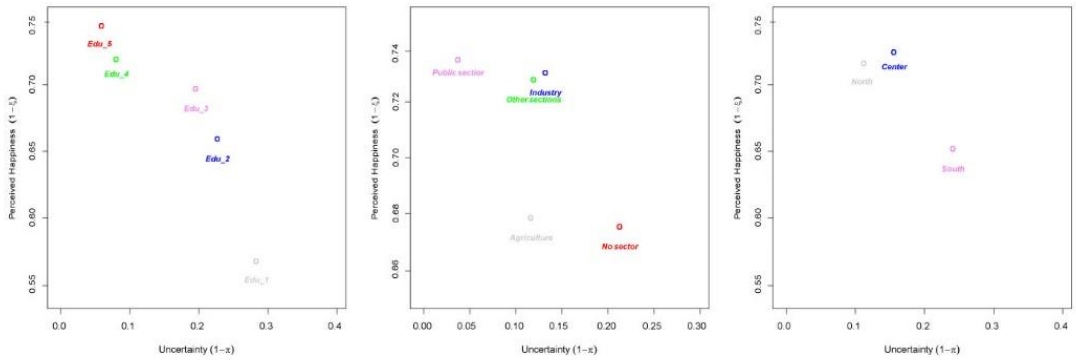
It is possible to observe a further simplification when we consider a nominal covariate as *marital status* (third panel). For a fixed level of uncertainty ($1 - \pi = 0.156$), different probability distributions are summarized by four points in the parameter space (Figure 2, forth panel). It turns out that single are happier than the others.

Fig. 2 CUB models for *perceived happiness* vs *gender* and *marital status*



In addition, the *perceived happiness* for subsample of respondents related to *education*, *sector of activity* and *geographical area* may be easily represented. In Figure 3 it is possible to observe higher scores for higher educated interviewees (first panel), who work in public sector (second panel) and live in the Center of Italy (third panel).

Fig. 3 CUB models for *perceived happiness* vs nominal/factor covariates (*education, sector of activity, geographical area*).



As an instance of continuous covariates we select *age* and *income*. Figure 4 (first panel) concerns a model in which the *age* covariate (more specifically, $age_i = \log(age_i) - \overline{\log(age_i)}$, for $i = 1, 2, \ldots, n$) has an impact on both parameters.

Also age^2, a parabolic effect, turns out to be significant on the feeling. By increasing the *age* of the respondent the level of satisfaction reduces. A peculiar feature of this approach is the simultaneous visualization of both effects in the parameter space (Figure 4, first panel) by varying the *age* of the respondents: young are happier than elderly people and are more resolute in their responses.

Income is another significant covariate for both components (second panel, Figure 4): the *perceived happiness* decreases with lower *income* whereas the level of uncertainty increases. The behaviour of the *income* covariate mimics an ordinal variable concerning the answer on the household income; specifically, survey asks if respondents consider sufficient to see the family through to the end of the month: this covariate is named *family condition* and ranges from 1 (with great difficulty) to 6 (very easily). The third panel of Figure 4 underlines the negative perception for people who express a lower expectation about this covariate.

The changing levels of *perceived happiness* may be shown in the same parameter space if we compare responses for the waves: 2008, 2010, 2012 (Figure 5, first panel). A higher *perceived happiness* in 2012 with respect to the 2010 wave (characterized by a higher uncertainty in the responses) is observed.

Another possible representation is to create some profiles of respondents for the analysis of *perceived happiness* as reported in Table 2 (the model concerns the 2012 wave).

In this more complex model, a significant impact of *education* and *family condition* for uncertainty, and of *gender* and *age* for feeling has been found. For an average *age* of 59 years the *perceived happiness* increases for higher educated women whereas the level of uncertainty increases for lower level of *education* and perceived *family condition*. The right panel of Figure 5 visualizes two selected profiles and shows how the effect of covariates appreciably changes the expected distribution of the responses.

Fig. 4 CUB models for *perceived happiness* vs continuous (first and second panel) and ordinal covariates (third panel)

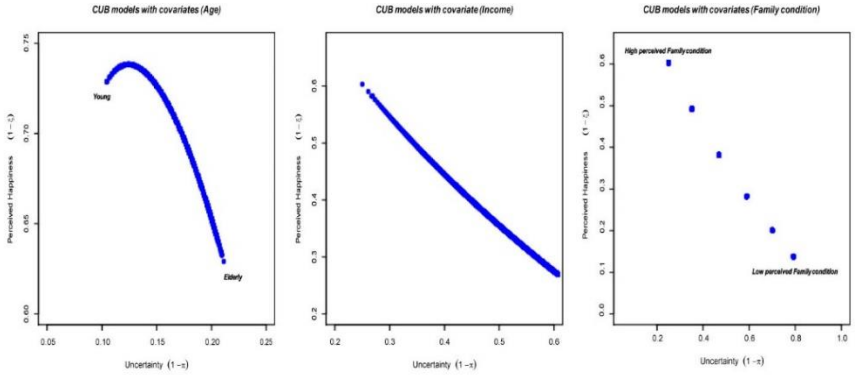


Table 2. CUB model for *Perceived Happiness* (wave 2012)

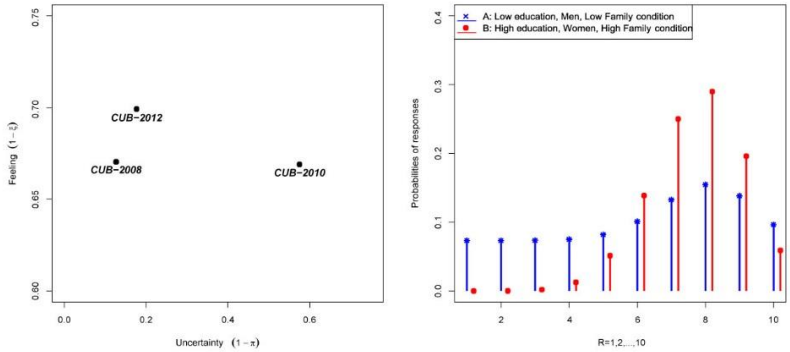
Components	Covariates	CUB model parameters	Wald-test
Uncertainty	constant	$\hat{\beta}_0 = -2.710(0.263)$	−10.291
	education	$\hat{\beta}_1 = 1.373(0.075)$	18.234
	family condition	$\hat{\beta}_2 = 0.323\ (0.074)$	4.467
Feeling	constant	$\hat{\gamma}_0 = -1.162\ (0.0$	−37.093
	gender	$\hat{\gamma}_1 = 0.167\ (0.020)$	8.203
	age	$\hat{\gamma}_2 = 0.417\ (0.037)$	11.301

3.2 *Perceived political trust* from European Social Survey

Data stem from European Social Survey (the ESS-Round 4 Project) available on www.europeansocialsurvey.org. It is an academically-driven social survey designed to visualize and explain the interaction between Europe's changing institutions and attitudes, beliefs and behaviour patterns of populations. The survey covers more than 30 nations and employs rigorous sample methodologies. Aims of this analysis is to explain the *perceived political trust* by means of the analysis of citizenship, involvement and democracy of different European Countries.

We analyse the perceived political action and the beliefs in Government projects by means of the *contextual* approach of CUB models. For each country (interpreted as a contextual covariate), the covariates we found significant to explain the perception of *political trust* are *education* (as an individual effect covariate) and *Gross Domestic Product* (*GDP*).

Fig. 5 CUB models for *perceived happiness* in three waves 2008, 2010, 2012 (left panel), and with respect to two specified profiles (right panel).



For m = 11 categories (from low to high political trust), a subsample size of n = 25, 000 citizens has been selected by a random draw from the whole sample of more than 35, 000 units from 21 countries. The estimated CUB model for the perceived political trust is reported in Table 3. A moderate level of uncertainty in the responses has been found since $1 - \hat{\pi} = 0.374$.

Table 3. CUB model with contextual effect for the perceived *Political trust*.

Components	Covariates	CUB model parameters	Wald-test
Uncertainty	constant	$\hat{\pi} = 0.626$ *(0.006)*	99.72
Feeling	constant	$\hat{\gamma}_0 = 0.269$ *(0.029)*	9.14
	education	$\hat{\gamma}_1 = -0.036$ *(0.002)*	−18.52
	GDP	$\hat{\gamma}_2 = 0.120$ *(0.004)*	26.35

This model predicts that the expected perception increases with the level of *education* of the *i*-th subject, and reduces with the *GDP* of the *j*-th country. It should also be possible to expand the inference in a *hierarchical* framework by considering mixed effects.

4. Conclusions

We start from the idea that methods for analysing a large amount of ordinal data (concerning the latent components of human well-being stemming from different sources in the context of official statistics) are a useful contribution to economic and social research. We have presented a framework which mimics the data generating process obtained by means of the selection of a category in a sequence of ordinal data. We have shown the relevant features of this class of models in terms of visualization and communicating statistics. The new approach for the analysis of collected data presents high flexibility and more parsimony with respect to the standard models.

The extended class suggests further studies aimed at analysing operational tools which regard the development of the European Statistical System towards 2020. The ability to summarize thousands of responses in a parameter space by means of an immediate idea of comparative feeling and uncertainty of several countries simplifies the presentation of the results. Finally, the philosophy of the approach is that data are used to derive the whole probability distribution of expected results. It simplifies the understanding of human behaviour when faced to a questionnaire or an interview in an effective way.

References

1. Corduas, M. **Analyzing bivariate ordinal data with CUB margins**, Statistical Modelling, Vo. 15, 2014, pp. 411-432
2. Faiella, I., Gambacorta, R., Iezzi, S. and Neri, A. **Italian Household Budgets in 2006**, Supplements to the Statistical Bulletin (new series), Vol. 7, 2008
3. Frey, B. S. and Stutzer, A. **Happiness and economics: how the economy and institutions affect well-being**, Priceton University Press, Priceton, 2002
4. Gambacorta, R. and Iannario, M. **Measuring job satisfaction with CUB models**, Labour, Vol. 27, 2013, pp. 198–224

5. Gambacorta, R., Iannario, M. and Valliant, R. **Design-based inference in a mixture model for ordinal variables for a stage stratified design**, Australian & New Zealand Journal of Statistics, Vol. 56, No. 2, 2014, pp. 125–143
6. Gottard, A., Iannario, M. and Piccolo, D. **Varying uncertainty in CUB models**. **Advances in Data Analysis and Classification**, DOI 10.1007/s11634-016-0235-0, 2016
7. Iannario, M, **On the identifiability of a mixture model for ordinal data**, Metron, Vol. LXVIII, 2010, pp. 87–94
8. Iannario, M. **Modelling shelter choices in a class of mixture models for ordinal responses**, Statistical Methods and Applications, Vol. 21, 2012, pp. 1–22
9. Iannario, M. **Hierarchical CUB Models for ordinal variables**, Communications in Statistics. Theory and Methods, Vol. 41, 2012, pp. 3110–3125
10. Iannario, M. **Modelling Uncertainty and Overdispersion in Ordinal Data**, Communications in Statistics. Theory and Methods, Vol. 43, 2014, pp. 771–786
11. Iannario, M. **Detecting latent components in ordinal data with overdispersion by means of a mixture distribution**, Quality & Quantity, Vol. 49, 2015, pp. 977–987
12. Iannario, M. and Piccolo, D. **CUB models: Statistical methods and empirical evidence**, in: Kenett, R.S. and Salini, S. (eds.) "Modern Analysis of Customer Surveys: with applications using R", Chichester: J. Wiley & Sons, 2012, pp. 231–258
13. Iannario, M. and Piccolo, D. **A comparative analysis of alternative frameworks for modelling ordinal data**, Working paper, 2014
14. Iannario, M. and Piccolo, D. **Inference for CUB models: a program,** R. Statistica & Applicazioni, Vol. XII, 2014, pp. 177-204
15. Iannario, M. and Piccolo, D. **A Generalized Framework for Modelling Ordinal Data**, Statistical Methods and Applications, 2015, DOI 10.1007/s10260-015-0316-9
16. Iannario, M. and Piccolo, D. **CUB: a class of mixture models for ordinal data. R package version 0.0.539,** 2015, http://CRAN.R-project.org/package=CUB
17. McCullagh, P. and Nelder, J. A. **Generalized linear models**. 2nd edition, Chapman and Hall, London, 1989
18. Piccolo, D. **On the moments of a mixture of uniform and shifted binomial random variables**, Quaderni di Statistica, Vol. 5, 2003, pp. 85–104
19. Piccolo, D. **Observed information matrix for MUB models**, Quaderni di Statistica, Vol. 8, 2006, pp. 33–78
20. Piccolo, D. and D'Elia, A. **A new approach for modelling consumers' preferences**, Food Quality and Preference, Vol. 19, 2008, pp. 247–259
21. Stiglitz, J., Sen, A. and Fitoussi, J. P. **Report of the Commission on the Measurement of Economic Performance and Social Progress**, Discussion paper, CMEPSP, 2009

[1] **Acknowledgements**
The authors gratefully acknowledge Editor and referees. This research has been supported by the SHAPE project within the frame of STAR Programme (CUP E68C13000020003) at University of Naples Federico II and by FIRB 2012 project (code RBFR12SHVV) at University of Perugia.

[2] Associate Professor in Statistics at University of Naples Federico II. She earned the PhD in 2005. She has been Visiting Researcher at Iowa University (USA), the University of Lancaster (UK), the University of Geneva (CH) and at Le Centre Européen des Sciences et du Goût (Dijone, France). She received the Fulbright Research scholarship for 2012. She collaborated to several national and international projects concerning methodological and applied re-

searches on mixture models for ordinal data a related estimation methods. Among them she is the Principal Investigator of the "Statistical Modelling of Human Perception" – Programme STAR. Her publication appeared on refereed journals and she has been invited to give some talks and seminars in several Universities and Center of Research. Recently, she performed a joint cooperation with the Bank of Italy with reference to modelling job satisfaction in a complex sample design.

[3] Full Professor of Statistics at Department of Political Sciences, University of Naples Federico II, since 1986. Senior Researcher of the Centro di Ricerche (Portici) from 1970 to 1986. Member of the Evaluation Board of University of Naples Federico II (2003 – 2008). Elected member of National Council of Italian Statistical Society. Chief of the projects VER01 and VER02: "Monitoring and Evaluating Orientation services", co-funded by the EU (FSE). Responsible of several research projects financed by MURST and CNR. Scientific Director of DESEC Project (1982-85) and ISTAT SARA Project (1997-98) which led to select a seasonal adjustment procedure for Italy. Visiting Fellow at Lancaster (UK), Madison (USA), Genève (CH), Madrid (E), Christchurch (NZ) Universities. Author of several papers on statistical inference, time series analysis, models for ordinal data. He is also author of some popular textbooks