A program in R for CUB models inference

ARTICLE · JANUARY 2009

2 AUTHORS, INCLUDING:

Domenico Piccolo
University of Naples Federico II
60 PUBLICATIONS **459** CITATIONS

SEE PROFILE

Version **2.0**, *www.dipstat.unina.it, September 2009*

A program in R for CUB models inference

Maria Iannario and Domenico Piccolo
Dipartimento di Scienze Statistiche, Università di Napoli Federico II
E-mail: {maria.iannario, domenico.piccolo}@unina.it

Summary: In this paper, we present the essential elements for using a program (version 2.0) implemented in the R statistical environment for CUB models inference. After reviewing basic definitions and notations, we discuss the genesis and usage of this class of models and their main inferential issues. Then, perspectives for future developments and an updated bibliography on this topic conclude the paper.

Keywords: Ordinal data modelling, CUB models, Shelter effect, R statistical environment.

1. Introduction

Starting from the beginning of 2000, a new approach for modelling discrete choices has been developed by statisticians working at Department of Statistical Science, University of Naples Federico II. The main characterization of the project –promoted by Domenico Piccolo– has been the awareness that selections and preferences are determined by psychological mechanisms which this class of models aims at explaining by using a parsimonious parametrization.

In this paper, we present a class of mixture models with covariates (hereafter, defined CUB models) which examines and compares the uncertainty of the answers and the feeling/adversion of the respondents towards the items, with the inclusion of subjects'/objects' covariates. Then, we extend this structure by taking the possible presence of a *shelter effect* into account.

More specifically, the purpose of this work is to introduce a program for CUB models inference and to describe the practical benefits of using this modelling tools for ordinal data specification, estimation, testing and interpretation. This software product is regularly updated by sequentially including new parts and improving existing ones. The current version (2.0) is a major upgrade of the previous ones since it increases computational efficiency, adds new information and indices to output and implements extended CUB models. Operationally, the main functions have been simplified and generalized.

Statistical foundations, inferential results and computational aspects of this program are not discussed in the present paper and more technical details are available in the references.

The work is organized as follows: in the next section, we establish notations for CUB probability structure (considerable attention is paid to notation), and in section

3 we briefly mention some inferential issues. In section 4 we extend the class of CUB models (*shelter choices*). Then, section 5 introduces CUB models in the R environment and review several information about the input/output processing steps. In section 6 we present instructions for plotting several estimated CUB models into the parametric space for comparison and interpretation. This is a remarkable feature of CUB modelling approach as it allows a clear visualization of hundreds or thousands of qualitative information in an appealing framework which is effective for classification, discrimination and detecting anomalous situations. In section 7, we introduce some new functions (implemented in the present version of the program) that are useful as graphical tools for more advanced statistical inference of CUB models. Some concluding remarks, a list of possible future developments and specialized references end the paper.

2. Specifications of CUB *models*

In statistical surveys, people are often asked to express judgements or evaluations on several topics; sometimes, it is required to make an ordered selection/arrangement in a definite list of m objects (items, services, sentences, etc.). A consistent hypothesis is that choices may depend both on raters and/or items. However, in order to simplify the analysis, we will assume hereafter that results are only related to subjects' characteristics (measures, opinions, status, and so on).

When we analyze ordered responses we have to distinguish between the expression of an evaluation about some fixed item (*rating*) and the assignment of a rank (*ranking*) to a given list of objects. For a given m, in the *rating* approach the answer of the subject is a single number for a fixed item; in the *ranking* approach, instead, the answer is a permutation of the first m integers, that is a vector of numbers specifying the sequential degree of preferences of the m objects.

For a correct understanding of CUB models, it is important to underline that our approach consists in the analysis of a univariate random variable related to the evaluation of the item (*rating*) or to the ordering of a single object (*ranking*). Notice that while in the first approach we study the univariate response of a group of subjects, in the second case CUB models are applied to a marginal analysis of ordinal data available as discrete multivariate random variables and whose components explain the stated preferences towards m fixed objects.

Of course, it is interesting to investigate logical and statistical relationships among this point of view and the common framework of literature, as it happens for Item Response Theory and Ordered Logit/Probit models, for instance: these lines of researches are currently under investigations[1]. However, we limit ourselves to describe this new approach as it has been consistently verified and successfully applied in several contexts.

As a matter of fact, CUB models explain, fit, and forecast the probability $Pr(R = r)$ that a discrete random variable R assumes values $r = 1, 2, \ldots, m$, for a given inte-

[1] We report that some preliminary results about the comparison between proportional odds and CUB models performances are encouraging in terms of predicting ability.

ger $m > 3$. They are based on a class of discrete probability distributions, originally called MUB, introduced by Piccolo (2003) and firstly applied by D'Elia and Piccolo (2005a, 2005b) with special reference to ranking data sets. These models have been generalized with the inclusion of different links, structures and covariates by a formal definition of CUB models (Piccolo, 2006). Recently, Iannario (2009a) proved that these models are identifiable and introduced a *shelter effect* by defining extended CUB models (Iannario, 2009c; Corduas et al., 2009). Instead, asymptotic standard errors of maximum likelihood estimates and related inferential issues are obtained according to results and notations as in Piccolo (2006), with some minor corrections.

Thus far, CUB models have been successfully applied in several fields including Linguistics (Balirano and Corduas, 2006, 2008; Cappelli and D'Elia, 2004, 2006b), Risk analysis (Cerchiello et al., 2009; Corduas, 2008c), Marketing (Cicia et al., 2008; Iannario and Piccolo, 2010; Piccolo et al., 2009), Teaching evaluations (Corduas, 2008b; D'Elia and Piccolo, 2006), University services performance (Cappelli and D'Elia, 2006b; Corduas et al., 2009; Iannario and Piccolo, 2009a), Measurement system analysis (Deldossi and Zappa, 2009), Medicine (D'Elia, 2008), Ranking preference (D'Elia and Piccolo, 2005a), Sociology (D'Elia and Piccolo, 2005b; Iannario, 2007a; 2008), Tourism (Iannario, 2009f), Subjective perception studies (Caliendo and Iannario, 2009; Iannario and Piccolo, 2009b, 2010b), Sensometrics (Piccolo and D'Elia, 2008), Qualitative analysis (Piccolo, 2008; Piccolo and Iannario, 2008b).

In these models, we assume that the final outcome of a process of judgement be a discrete observation generated by an investigated trait which is intrinsically continuous. More specifically, when people are faced with discrete choices, the psychological mechanism, by which the elicitation is accomplished, is the result of a personal *feeling* towards the object and an inherent fuzziness/*uncertainty* in choosing the ordinal value of the response. *Feeling* is usually related to the subjects' motivations whereas *uncertainty* mostly depends by circumstances that surround the elicitation process.

In this regards, *shifted Binomial* random variable is an adequate probability model for representing the discrete version of a latent judgement process, by mapping a continuous and unobserved evaluation into a discrete set of values belonging to $\{1, 2, \ldots, m\}$. This random variable shows high flexibility with respect to the location of a mode and skewness and simple generalizations of this structure allow further options.

On the other side, the *discrete Uniform* random variable is a suitable structure for describing the inherent uncertainty of a discrete choice process, since it represents the model with maximum entropy on a finite discrete support. Thus, any observed uncertainty contained in the data may be weighted with respect to this extreme case.

On this basis, we consider r (=the observed ordered measure) as the realization of a discrete random variable R defined as a mixture of a Uniform and a Shifted Binomial random variables. Formally, its probability mass function is defined by:

$$P_r\left(R = r\right) = \pi \binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r} + (1-\pi)\,\frac{1}{m}\,, \quad r = 1, 2, \ldots, m\,,$$

with $\pi \in (0, 1]$ and $\xi \in [0, 1]$.

Expectation and variance of R are given by:

$$E(R) = \pi\ (m-1)\left(\frac{1}{2} - \xi\right) + \frac{(m+1)}{2};$$

$$Var(R) = (m-1)\left\{\pi\,\xi\,(1-\xi) + (1-\pi)\left[\frac{m+1}{12} + \pi\,(m-1)\left(\frac{1}{2} - \xi\right)^2\right]\right\}.$$

It is immediate to realize that π is a parameter inversely related to the weight of the uncertainty component; thus, $(1-\pi)/m$ is a measure of the *uncertainty share* which spreads uniformly over the support. Instead, the interpretation of ξ changes with the setting of the analysis since it depends on how the responses have been coded (the first position represents the higher feeling/concern and the last one the lower, or vice versa). According to the context, ξ has been interpreted as *degree of perception*, *index of selectiveness/awareness*, *measure of concern*, *threshold of pain*, *subjective probability*, and so on.

Better solutions are usually obtained when we introduce the subjects' *covariates* aimed at relating both the feeling and the uncertainty to the respondents' features. In fact, the acronym CUB is originated by the presence of ***C***ovariates in the mixture of ***U***niform and shifted ***B***inomial distributions.

A peculiar aspect of these models is the direct relationship among parameters and covariates, without using expectations –as it is standard in the Generalized Linear models (GLM) paradigm– by means of a monotone function (logistic, in several cases).

If covariates are significant, they improve model fitting and allow for better discrimination among different sub-populations (for instance, via dummies covariates, as in Iannario, 2007b, or by clustering methods as in Corduas, 2008a,b,c). Actually, also objects' covariates may be introduced but these aspects will not be discussed here (Piccolo and D'Elia, 2008).

In order to set a standard for the notation, hereafter, we will classify CUB models following the scheme:

Models	*Covariates*	*Parameter vectors*	*Parameter spaces*	*Number of parameters*
CUB $(0,0)$	no covariates	$\boldsymbol{\theta} = (\pi, \xi)'$	$(0,1] \times [0,1]$	2
CUB $(p,0)$	covariates for π	$\boldsymbol{\theta} = (\boldsymbol{\beta}', \xi)'$	$R^{p+1} \times [0,1]$	$p+2$
CUB $(0,q)$	covariates for ξ	$\boldsymbol{\theta} = (\pi, \boldsymbol{\gamma}')'$	$(0,1] \times R^{q+1}$	$q+2$
CUB (p,q)	covariates for π and ξ	$\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$	R^{p+q+2}	$p+q+2$

Then, the general formulation of a CUB (p,q) model (with p covariates to explain uncertainty and q covariates to explain feeling) is expressed by:

1. A *stochastic component*:

$$Pr(R_i = r \mid \boldsymbol{y}_i;\, \boldsymbol{w}_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} + (1-\pi_i)\left(\frac{1}{m}\right),$$

for $r = 1, 2, \ldots, m$, and for any i-th subject, $i = 1, 2, \ldots, n$.

2. Two *systematic components*:

$$\pi_i = \frac{1}{1 + e^{-\boldsymbol{y}_i \boldsymbol{\beta}}}; \quad \xi_i = \frac{1}{1 + e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}}; \quad i = 1, 2, \ldots, n;$$

where $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the subjects' covariates for explaining π_i e ξ_i, respectively.

With respect to classical GLM approach to ordinal data, CUB models offer a straightforward relationship between a probability statement for ordinal answers and subjects' covariates. Moreover, although latent variables are conceptually necessary in order to specify the nature of the mixture components, the inferential procedures are not based upon the knowledge (or estimation) of cutpoints. As a consequence, if the model is adequate, this simplification turns out to be a more parsimonious parameterization.

3. Inferential issues in CUB *models*

Given a sample of observed values of ratings and covariates $(r_i, \boldsymbol{y}_i, \boldsymbol{w}_i)'$, for $i = 1, 2, \ldots, n$, the log-likelihood function for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ is defined by:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\frac{1}{1 + e^{-\boldsymbol{y}_i \boldsymbol{\beta}}} \left\{ \binom{m-1}{r_i - 1} \frac{e^{(-\boldsymbol{w}_i \boldsymbol{\gamma})(r_i - 1)}}{(1 + e^{-\boldsymbol{w}_i \boldsymbol{\gamma}})^{m-1}} - \frac{1}{m} \right\} + \frac{1}{m} \right].$$

Inferential issues for the joint efficient estimation of the parameters are discussed in details by Piccolo (2006) who derived an EM algorithm for the maximum likelihood estimation of the parameter vector $\boldsymbol{\theta}$. Moreover, several proposals for improving initial estimates of this procedure are now included in version 2.0.

Then, in order to assess the significance of the estimated parameters, and the relevance of the covariates for explaining the main features of the data, we rely on the asymptotic inference of maximum likelihood estimators. Thus, we compare log-likelihoods of CUB models by means of their corresponding deviances difference, as in the following scheme:

Comparisons	*Deviances difference*	*Degrees of freedom*
$CUB(p, 0)$ *versus* $CUB(0, 0)$	$2\ (\ell_{10} - \ell_{00})$	p
$CUB(0, q)$ *versus* $CUB(0, 0)$	$2\ (\ell_{01} - \ell_{00})$	q
$CUB(p, q)$ *versus* $CUB(0, 0)$	$2\ (\ell_{11} - \ell_{00})$	$p + q$

For a given $m > 3$, information contained in the ordinal data $(r_1, r_2, \ldots, r_n)'$ are strictly equivalent to that contained in the frequencies $(n_1, n_2, \ldots, n_m)'$ of the ordered categories. Thus, the log-likelihood for the *saturated* CUB model is obtained:

$$\ell_{sat} = -n\ \log(n) + \sum_{r=1}^{m} n_r\ \log(n_r);$$

For judging the goodness-of-fit of the estimated model, several measures may be derived. Specifically, a normalized dissimilarity index $(= Diss)$ is defined as the absolute distance among the observed relative frequencies f_r and the probabilities $p_r(\hat{\boldsymbol{\theta}}) = P_r\left(R = r|\hat{\boldsymbol{\theta}}\right)$ estimated by the model. A further index is $ICON$ (=*I*nformation *CON*tent), that is a pseudo-R^2; it measures the improvement we obtain when we move from a completely uninformative distribution (as the Uniform one) to a well structured random variable (as CUB models), without or with covariates. In order to compare different models, we prefer to consider BIC index. Then, in version 2.0, a new *likelihood-based fitting measure*, denoted by D^2, has been proposed and we are currently testing its efficacy.

The formulas for the above mentioned measures are:

$$Diss = \frac{1}{2}\sum_{r=1}^{m}\left|f_r - p_r(\hat{\boldsymbol{\theta}})\right|; \qquad ICON = 1 + \frac{\ell(\hat{\boldsymbol{\theta}})/n}{\log(m)};$$

$$D^2 = \frac{1}{m}\sum_{r=1}^{m}\left(\frac{f_r}{p_r(\hat{\boldsymbol{\theta}})}\right)^2 - 1; \qquad BIC = -2\,\ell\left(\hat{\boldsymbol{\theta}}\right) + p\,\log(n).$$

Of course, all these measures (except $ICON$) should be as minimum as possible for a good fitting result.

4. Extended CUB *models*

A noticeable improvement of version 2.0 is the implementation of an extended version of CUB models, where we add a third component for explaining a (possible) specific behavior of some respondents. This effect has been observed when people are asked to select a single option among a list of m ordered choices and a single modality receives a biased amount of preference due, for instance, to its position, linguistic expression, numbers roundedness, adverse/favorite selection with respect to a given standard and expected response model, and so on.

The circumstance that such models have been effective for explaining cases where people preferred a simplified category (that is, a *shelter choice*), in order to avoid more elaborate decisions, motivates its name (Iannario, 2009c). In addition, it should be noted that in some circumstances –if one misses this component– estimates are biased and inefficient, and fitting and predictions are not satisfactory.

Specifically, an *extended* CUB *model* is defined for $r = 1, 2, \ldots, m$ by:

$$P_r\left(R = r|\boldsymbol{\theta}\right) = \pi_1\binom{m-1}{r-1}\xi^{m-r}(1-\xi)^{r-1} + \pi_2\,\frac{1}{m} + (1-\pi_1-\pi_2)\,D_r^{(c)},$$

where $\boldsymbol{\theta} = (\pi_1,\, \pi_2,\, \xi)'$ is the extended parameter vector and $D_r^{(c)}$ is a degenerate random variable whose probability mass is concentrated at $r = c$, that is:

$$D_r^{(c)} = \begin{cases} 1, & \text{if } r = c; \\ 0, & \text{otherwise.} \end{cases}$$

Notice that we are supposing that c is a known integer belonging to $\{1, 2, \ldots, m\}$. Extended CUB models are identifiable only for $m > 4$.

We should observe that if $\pi_1 + \pi_2 = 1$ the extended CUB model collapses to the standard one; thus, the presence of a possible *shelter choice* may be tested on sampled data by means of standard significance tests. In fact, the parameter $\delta = 1 - \pi_1 - \pi_2$ measures the added relative contribution of the *shelter choice* at $y = c$ with respect to the standard version of the model.

Finally, the extended model is able to fit distributions where most of the respondents' choices are concentrated at a single intermediate category: in this (rare) situation we will get: $\pi_1 = \pi_2 = 0$.

We are planning to release future versions of this program by generalizing the present formalization of CUB models by taking subjects' covariates into account also for explaining a possible *shelter effect*.

5. Handling input and output information

This section concerns the introduction of essential elements for using a CUB models program, implemented in the R statistical environment.

The current R code derives from previous releases originally obtained as GAUSS© program; thus, it is consistent with the logic of GAUSS procedures[2]. In version 2.0, there are *macro functions* which include micro functions as a nested system. Thus, the program is active by entering the `CUB()` function whose options call more specialized functions (related to different CUB models); at the lowest level, *local functions* perform more specific tasks.

As already introduced, some aspects are worth of consideration. Ordinal responses are usually considered as rating in $1, 2, ..., m$, with $m > 3$, but it is also possible to model ranking data as univariate marginal distributions. Thanks to the *reversibility* of CUB models, we may change the scale of rating/ranking (if required) by putting: `ordinal=m-ordinal+1`. Anyway, m is a global variable which must be declared before calling any function of the program.

We denote by `filename.txt` the file containing the data and by `ordinal` the vector of ordinal data (which are the object of the analysis); then, $\boldsymbol{Y}$ and $\boldsymbol{W}$ are the covariates matrices for explaining π and ξ, respectively, when necessary. These data should be available as a text (or R) file in standard matrix format (with columns names in the first row). Notice that these matrices *do not contain* 1's in the first column.

Thus, we will read data and set the number of categories:

```
> dataset=read.table("filename.txt",header=T)}
> m=number_of_categories
```

Then, the program is active in R by means of:

```
> source("CUB.R")
```

[2] GAUSS programs have been written by Domenico Piccolo during $2001 - 2005$ and some conventions have been imported also in R. We found that these conventions make more readable the listing of the program, as advocated in the books by Verzani (2005) and Spector (2008).

Then, the instruction for estimation, testing and inference is:

```
> CUB(ordinal,Y=paicov, W=csicov, shelter=c)
```

Default values are:

```
> CUB(ordinal,Y=0, W=0, shelter=0)
```

As a consequence, for inferring on a CUB model without covariates is sufficient to run:

```
> CUB(ordinal)
```

In the program, we denote by `ordinal` the vector of rating data to be modelled, and by `paicov, csicov` the (possible) covariates, whereas `c` is the category where a *shelter* effect should be (possibly) checked. Table 1 summarizes instructions for getting different CUB models inference.

Table 1. Instructions for CUB *models estimation.*

Model	*Covariates*	*Instruction*
CUB $(0,0)$	$====$	`CUB(ordinal)`
CUB $(p,0)$	`paicov`	`CUB(ordinal,Y=paicov)`
CUB $(0,q)$	`csicov`	`CUB(ordinal,W=csicov)`
CUB (p,q)	`paicov,csicov`	`CUB(ordinal,Y=paicov,W=csicov)`
CUB $+shelter$	$====$	`CUB(ordinal,shelter=c)`

In some applications, only the aggregated frequencies $(n_1, n_2, ..., n_m)'$ of the responses are available, for instance in a vector `frequencies`. Since the program is able to process n distinct ratings we have to expand such information in order to generate n ordinal values. This is immediate by means of:

```
> ordinal=rep(1:m,frequencies)
> CUB(ordinal)
```

Notice that `frequencies` must be a vector of length m even if some observed frequencies are 0.

Output of the program varies in function of the presence/absence of covariates and of their nature (dichotomous or not). Specifically, in addition to common statistical outputs (estimates, standard errors, *p*-values, fitting measures), the program offers some graphical information according to the following criteria:

- CUB $(0,0)$ model: a plot of observed frequency distribution of ordinal categories and related estimated probabilities is shown.
- CUB $(1,0)$ model: *if and only if* a single covariate for π is a *dichotomous* one and *if it is coded* as $0, 1$, plots of estimated CUB distributions for the two subgroups are automatically generated, with several indicators for testing significance of their diversity.

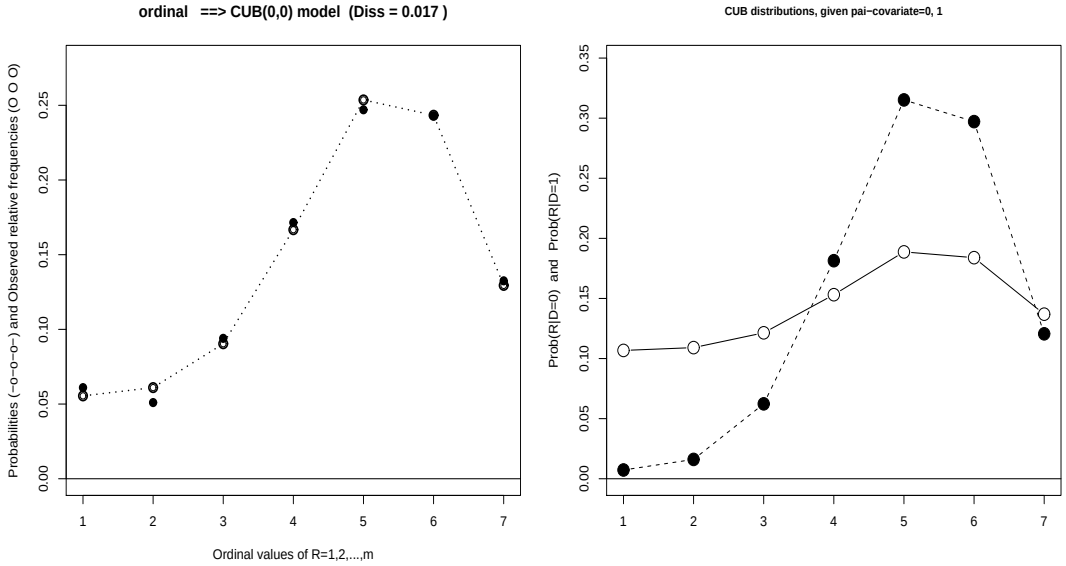

Figure 1. Estimated CUB *models without (left panel) and with a dummy covariate (right panel).*

- CUB $(0, 1)$ model: *if and only if* a single covariate for ξ is a *dichotomous* one and *if it is coded* as $0, 1$, plots of estimated CUB distributions for the two subgroups are automatically generated, with several indicators for testing significance of their diversity.
- CUB (p, q) model: currently, no plot is automatically generated. Users are encouraged to study expected responses as functions of covariates (if continuous) or conditioned (if dichotomous or politomous).
- CUB $+ \, shelter$ model: a plot of observed frequency distribution of ordinal categories and related estimated probabilities is shown.

Figure 1 shows standard plots of estimated CUB models without covariates and with a dummy covariates (for explaining a π significant behavior), respectively. In this regard, we observe plots of probability distributions are joined, although they concern a discrete support; in fact, we realized that this choice helps very much for detecting and comparing different patterns and shapes.

The following list is a standard output generated by running a CUB model without covariates.

```
=====================================================================
> CUB 2.0 ============================================================
=====================================================================
=====>>> C U B (0,0) model  <<<=====   ML-estimates via E-M algorithm
=====================================================================
```

```
*** m= 7   *** Sample size: n= 20184    *** Iterations= 39  (maxiter=250)
=========================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
=========================================================================
pai          0.86684          0.00459          188.85        0
csi          0.16287          0.00143          113.90        0
=========================================================================
Variance-covariance matrix
      [,1] [,2]
[1,] 2e-05    0
[2,] 0e+00    0
=========================================================================
Parameters correlation matrix
        [,1]    [,2]
[1,] 1.00000 0.39864
[2,] 0.39864 1.00000
=========================================================================
Log-lik(pai^,csi^) = -30383
Mean Log-likelihood= -1.5053
Log-lik(saturated) = -29952
Deviance           = 862.43
-------------------------------------------------------------------------
Log-lik(UNIFORM)         = -39276
Log-lik(Shifted-BINOMIAL)= -32413
-------------------------------------------------------------------------
AIC-CUB00          = 60770
BIC-CUB00          = 60785
ICON(CUB00)        = 0.22643
=========================================================================
Uncertainty Share: {(1-pai)/m} =  0.019022 ( 0.00065564 )
Asympt.Conf.Int.(95%)          =[ 0.017737 ;  0.020308 ]
=========================================================================
Pearson Fitting measure    ==>  X^2 = 906.26 (p-val.= 0 )
Lik-based fitting measure  ==>  D^2 = 0.079113
Normed Dissimilarity index ==>  Diss= 0.086476
=========================================================================
Observed average          = 5.7806  Sample variance        = 1.7015
Expectation of R~CUB(0,0) = 5.7534  Variance of R~CUB(0,0) = 1.7140
=========================================================================
Elapsed time= 0.05 seconds ===============>>> Sat Sep 05 19:51:21 2009
=========================================================================
```

Instead, the following list is a reduced output of a CUB $(3, 6)$ model fitted to a very large data set concerning perceptions. We observe that all covariates are significant (we retain β_1 since it improves likelihood) and also consistent with the expected sign and interpretation for this real case study.

```
=========================================================================
> CUB 2.0 ================================================================
=========================================================================
=====>>> C U B (p,q) model <<<=====   ML-estimates via E-M algorithm
=========================================================================
 Covariates for pai ==> p= 3      and     Covariates for csi ==> q= 6
```

```
=====================================================================
*** m= 7   *** Sample size: n= 20184    *** Iterations= 39  (maxiter=250)
=====================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
=====================================================================
beta_0        1.5105         0.07654         19.735        0
beta_1        0.03231        0.08103          0.39874        0.69008
beta_2       -1.3101         0.12569        -10.423        0
beta_3        2.534          0.32822          7.7204         1.1546e-14
gamma_0      -1.6028         0.02519        -63.628        0
gamma_1       0.10423        0.02184          4.7724         1.8201e-06
gamma_2       0.24576        0.06186          3.9728         7.1021e-05
gamma_3       0.28961        0.0908           3.1895         0.001425
gamma_4      -0.04701        0.02235         -2.1034         0.035434
gamma_5      -0.22604        0.0285          -7.9312         2.2205e-15
gamma_6      -0.55917        0.07409         -7.5472         4.4409e-14
=====================================================================
                          Variance-covariance matrix
.....................................................................
=====================================================================
                          Parameters correlation matrix
.....................................................................
=====================================================================
Log-lik(beta^,gamma^)= -30191
Mean Log-likelihood  = -1.4958
---------------------------------------------------------------------
AIC-CUBpq          = 60403
BIC-CUBpq          = 60490
ICON(CUBpq)        = 0.23133
=====================================================================
=====================================================================
Elapsed time= 96.18 seconds ===============>>> Sat Sep 05 20:08:02 2009
=====================================================================
```

After running the models, the program allows the possibility to use estimated or computed quantities, thanks to the assignment facilities of R. In this way, users may plot and/or save any useful values for comparing models characteristics obtained on different data sets, subgroups, time periods, and so on.

With obvious meaning, all available variables are shown in Table 2.

Table 2. List of available variables after running CUB *models.*

Model	*Variables*
CUB $(0,0)$	`pai,csi,varmat,loglik,diss,n,AICCUB00,BICCUB00`
CUB $(p,0)$	`bet,csi,varmat,loglik,diss,n,AICCUBp0,BICCUBp0`
CUB $(0,q)$	`pai,gama,varmat,loglik,diss,n,AICCUB0q,BICCUB0q`
CUB (p,q)	`bet,gama,varmat,loglik,diss,n,AICCUBpq,BICCUBpq`
CUB $+\ shelter$	`pai1,pai2,csi,varmat,loglik,diss`

6. Visualization of CUB models in the parametric space

A relevant feature of this approach is the possibility to plot estimated CUB models as single points into the parametric space and to interpret (with some caution) the characteristics of hundreds or even thousands of ordered data in terms of closeness, clustering and other useful issues. This opportunity has been successfully applied in several papers (see the references) and require some skill since sampling variability of π and ξ estimates are generally different. Indeed, Euclidean metric is not the best one to consider in this space and a naive criterion to assess distance among estimated models is not statistically effective. These issues are related to clustering and similarity of CUB models as fully discussed by Corduas (2008a,b,c) and deserve more investigations. In the following, we will limit ourselves to present a simple example in order to show how to get such results.

We suppose that the available ordered data are ratings about $k = 8$ items on a $m = 7$ points Likert scale. They are available in the ASCII file `dataset` and the relevant items of interest are denoted by `ITEM1, ITEM2, ..., ITEM8`, respectively; we suppose also that rating of the items are contained in the first eight columns of the data frame. The following codes read data, estimate 8 different CUB models, save the estimated parameters in the vectors `vettpai` and `vettcsi`, respectively; then, the estimated CUB models are plotted and labelled in the parameter space.

Moreover, by activating the library `ellipse` (an optional program of `R`), the final instructions plot also the asymptotic confidence ellipse for each item in order to visualize significant differences among them. This required to save also the variance-covariances matrices in a convenient list.

Of course, if ordinal data to be modelled are not contained in adjacent columns of the data set (or are available in different data sets), users of R should modify the proposed code. Finally, thanks to the function `sink()`, all output may be (optionally) directed to and saved into a file for future checks.

```
> source("CUBJ.R");
> m=7;
> dati=read.table(".....",header=TRUE);
> nameitems=c("ITEM1","ITEM2","ITEM3","ITEM4",
            "ITEM5","ITEM6","ITEM7","ITEM8");
> k=length(nameitems);
> vettpai=rep(NA,k); vettcsi=rep(NA,k); vmat=vector("list",k);

> for(j in 1:k){
      with(data=dati,CUBJ(dati[,j]));
      vettpai[j]=pai; vettcsi[j]=csi; vmat[[j]]=varmat;
  }

> plot(vettpai,vettcsi,xlim=c(0.7,1),ylim=c(0.21,0.32),
```

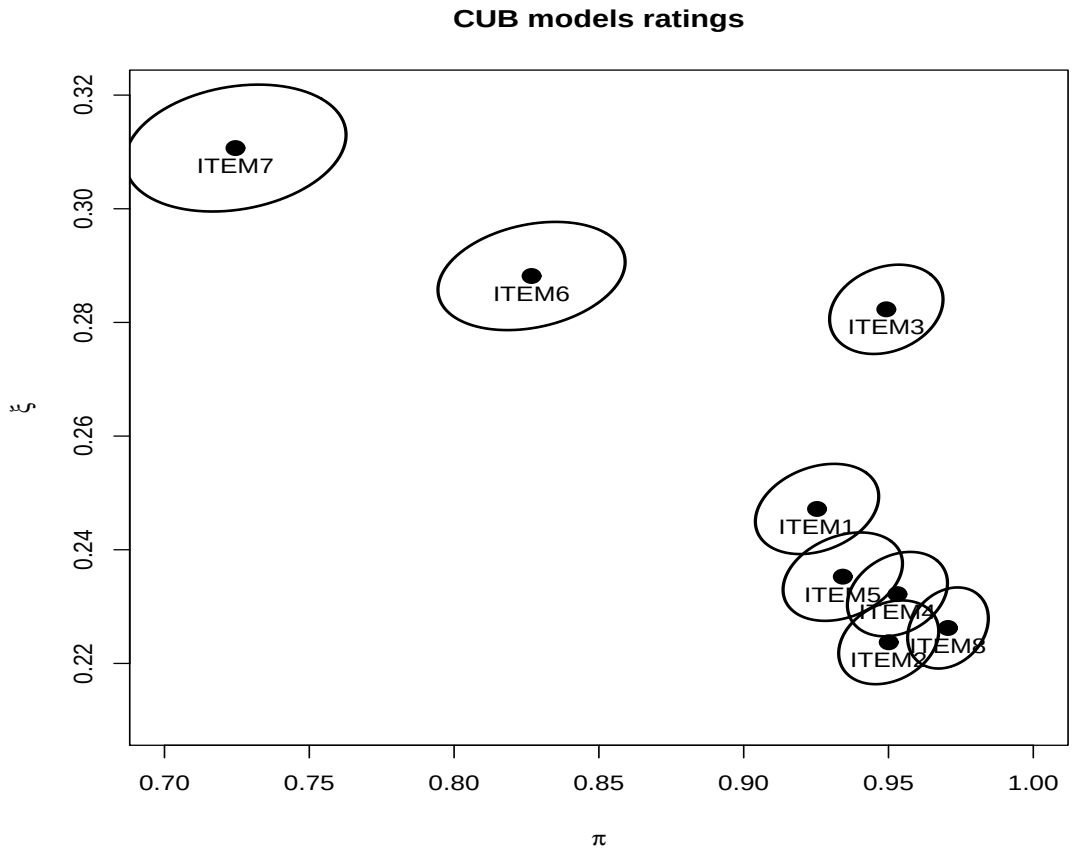

Figure 2. Visualization of estimated CUB *models in the parametric space.*

```
        xlab=expression(pi),ylab=expression(xi),
        main="CUB models ratings",pch=19,lwd=5);
> text(vettpai,vettcsi,labels=nameitems,pos=1);

> library(ellipse);
> for(jj in 1:k){
        lines(ellipse(vmat[[jj]],centre=c(vettpai[jj],vettcsi[jj])),
        lwd=2);
  }
```

In Figure 2 we show the graphical output derived from the previous code. In this data set the evaluations of thousands of people about 8 items have been sharply summarized by mapping estimated CUB models into the parametric space. Thus, the final results points out that items 3, 6 and 7 receive different feelings from users, also with increasing uncertainty. On the contrary, there is some overlapping among the remaining ones. Notice that we zoomed the relevant area of the parametric space in order to improve the visual detection of possible differences among the items.

7. Further tools for advanced analyses

In version 2.0, we are also delivering several functions for improving understanding and research on CUB models, as for instance those related to expectation and variance of CUB models. However, we quote two of them as they are relevant tools for obtaining more refined plots and effective simulation routines, respectively.

After running a CUB model, it is possible to exploit some *ad hoc* functions (released within the program) to deepen the inferential results. In fact, it is possible to obtain the *log-likelihood contours* plot and the *profile log-likelihood* functions[3] for both parameters of a CUB model without covariates. This may be accomplished by the following instructions:

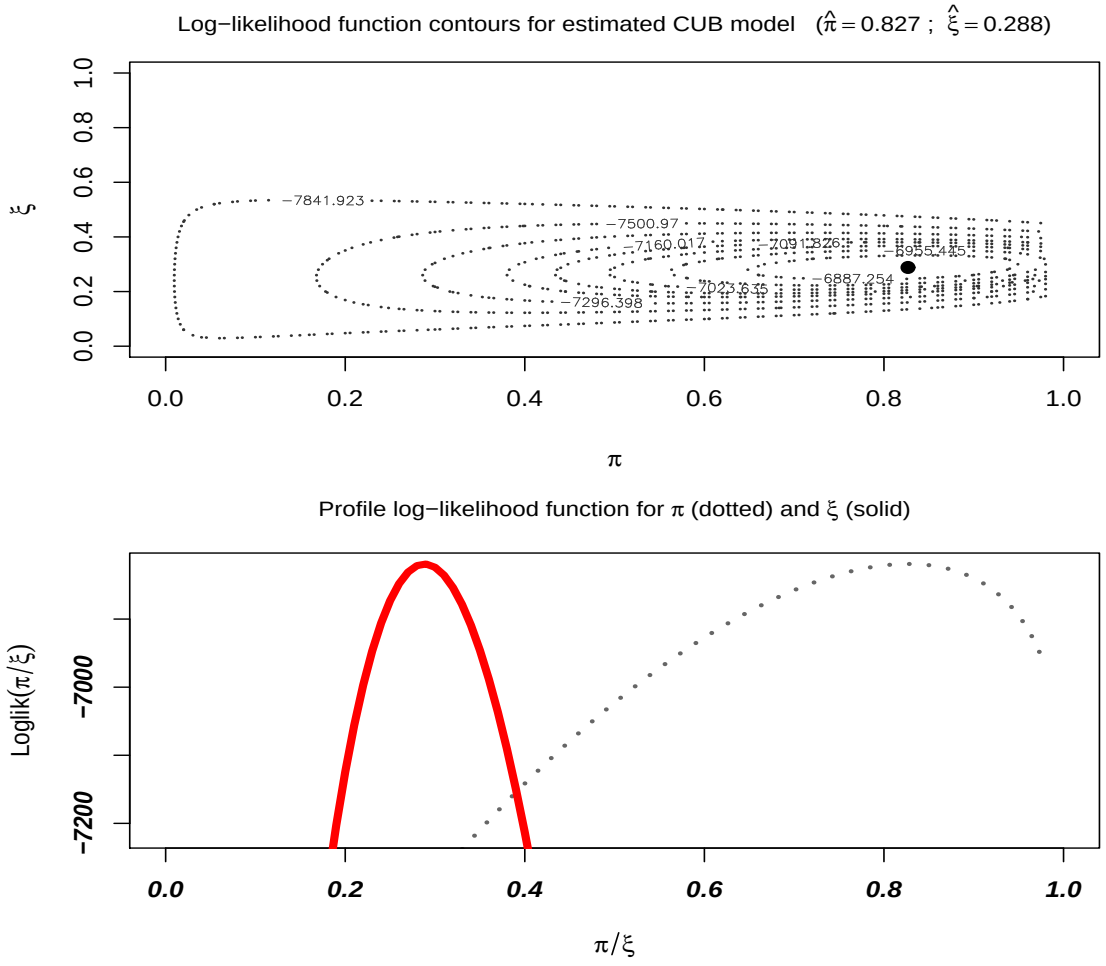

Figure 3. Contours plot and profile log-likelihood functions.

[3] In the code of the proposed functions, it is also available (but now it is commented out) a command for getting a 3D surface plot of the loglikelihood function. Moreover, profile loglikelihood function includes also a `ctune` option useful for a better display of these two functions on the same scale.

```
> source("CUB.R");
> m=.........                          # number of categories
> CUB(ordinal);
> par(mfrow=c(2,1));                   # set 2x1 graphical area
> par(mar=c(4,4,3,1)+0.1);             # set new margins
>    elleplot(pai,csi,loglik,ordinal); # contours plot
>    proflik(pai,csi,ordinal);         # profile functions
> par(mar=c(5,4,4,2)+0.1);             # reset standard margins
> par(mfrow=c(1,1));                   # reset 1x1 graphical area
```

Figure 3 shows a typical example of application of the previous instruction to a real data set. Notice how confidence intervals for uncertainty parameter (π) are larger than those for the feeling parameter (ξ), a situation quite common for real data set. This implies that vertical displacements are by far the most important for assessing significant distances in the parametric space.

We observe that such instructions must be executed *after* a CUB model has been estimated since these graphical functions require the parameters generated by the `CUB()` function.

Finally, for making experience or for doing research on CUB models, we propose a simulation routine called `simcub()`. It is active after calling the main `CUB()` function. Then, the following instructions will generate `nsimul` pseudo-random numbers from a CUB model with given m, π, ξ:

```
> source("CUB.R")
> simcub(nsimul,m,pai,csi)
```

We report that, on a plain PC with 3Gb of RAM, the elapsed time in seconds turns out to be an almost linear function of `nsimul`. In fact, after running several samples of increasing sizes, we get the following estimated regression:

$$\widehat{Time} = 0.063 * (nsimul/100000)\,.$$

The correlation coefficient among observed and predicted values is 0.999. Thus, for obtaining 10 millions of pseudo-random numbers we predict to wait for about 7 seconds.

8. Future developments

In this paper, we have briefly presented the main characteristics of CUB models for studying ordinal data by means of a related software (version 2.0) implemented in the `R` environment.

Some of the research lines we are currently exploring are the following:

- The current program requires some further refinements in order to be released with the logic of a standard `R` package.

- An extension of CUB models is necessary for achieving a multivariate approach, in order to analyze at the same time the whole responses to items.
- In several instances (educational evaluations, risk analyses, stratified consumers), data are collected on the basis of some hierarchical structure (schools, departments, income classes); then, our research aims at introducing a multilevel (hierarchical) approach in CUB models.
- In several situations, it may happen that raters belong to separated clusters, so that the effect of belonging to a group can modify the expressed rating. This is the case, for example, of longitudinal surveys when we must consider the correlation among the ranks expressed by the same rater during the time, or also the case of surveys on different groups of consumers with homogeneous behaviour.
- Data mining of ratings and preferences, especially in the case where panel of thousands of respondents regularly express their evaluations, is an innovative field where the sharp simplification and interpretation of CUB model would be worth of interest.
- Both uncertainty and feeling are personal latent traits upon which researchers of a definite area generally possesses long experience and a priori beliefs; as a consequences, a Bayesian approach to CUB models is worth of interest.
- Small sample size inference as well as the performance of parametric and non-parametric approaches (as bootstrap and permutation tests, for instance) are further topics to be investigated.
- From a computational point of view, multi-step procedures for getting effective and consistent acceleration of the EM procedure are objects of recent investigations.

`Disclaimer`: *This program is freely released without any responsibility for the Authors and cannot be distributed in part and/or without all comments included in the code. Users are requested to correctly quote any usage of this program in their work and kindly inform Authors by E-mail of any errors, bugs, suggestions, amendments, proposals, improvements. A pdf copy of any published paper and/or report where this program and/or* CUB *models are successfully applied is warmly requested. Finally, we are engaged to continuously improve and diffuse the current state of art in this area by maintaining this software on the Department website; moreover, we will record papers, data and applications which refer to* CUB *models by updating the list of references.*

Acknowledgements: The work has been partly supported by submitted PRIN-2008 research project: "Modelli per variabili latenti basati su dati ordinali: metodi statistici ed evidenze empiriche" and proposed FIRB project "Modelli statistici di analisi di dati per i servizi accademici"; moreover, it

has benefitted from research structures of CFEPSR, Portici. We acknowledge several suggestions and encouragements received by many colleagues that helped us in the revision and improvements of preliminary versions.

Selected references about CUB *models*

Balirano G., Corduas M. (2006) Statistical methods for the linguistic analysis of a humorous TV sketch show, *Quaderni di Statistica*, 8, 101–124.

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

Caliendo G., Iannario M. (2009) Communicating European values in institutional discourse: a statistical model for the analysis of citizens' perception of the EU, submitted for publication.

Cappelli C., D'Elia A. (2004) La percezione della sinonimia: un'analisi statistica mediante modelli per ranghi, in: *Le poids des mots - Actes de JADT2004* (Prunelle G., Fairon C. and Dister A. Eds.), Presses Universitaires de Louvain, Belgium, 229–240.

Cappelli C., D'Elia A. (2006a) Preliminary identification of relevant covariates in models for ordinal data: a strategy based on tree methods, in: (Croux C., Riani M., Chiandotto B., Bini M. and Bertaccini B., Editors) *ROBCLA-06 Book of Abstracts*, Università di Firenze, 23–24.

Cappelli C., D'Elia A. (2006b) A tree-based method for variable selection in models for ordinal data, *Quaderni di Statistica*, 8, 125–135.

Cerchiello P., Iannario M., Piccolo D. (2008) Assessing risk perception by means of ordinal models, *Proceeding of 2008 MAF Meeting*, Springer, 65–73.

Cicia G., Corduas M., Del Giudice T., Piccolo D. (2008) L'analisi delle preferenze dei consumatori nei confronti delle produzioni di qualità: uno studio del consumo di caffè equosolidale mediante il modello CUB , SIDEA Meeting, Portici, 25-27 September 2008.

Corduas M. (2008a) Clustering CUB models by Kullback-Liebler divergence, *Proceedings of SCF-CLAFAG Meeting*, ESI, Napoli, 245–248.

Corduas M. (2008b) A study on University students' opinions about teaching quality: a model based approach for clustering data, *Proceedings of DIVAGO Meeting*, University of Palermo, Springer, forthcoming.

Corduas M. (2008c) Statistical procedures for clustering ordinal data, *Quaderni di Statistica*, 10, 177–189.

Corduas M., Iannario M., Piccolo D. (2009) A class of statistical models for evaluating services and performances, in: M.Bini, P.Monari, D.Piccolo, L.Salmaso (eds.): *Statistical methods for the evaluation of educational services and quality of products*, Contribution to Statistics, Springer, 99–117.

Deldossi L., Zappa D. (2009) Measurement errors and uncertainty:a statistical perspective, in: S.Ingrassia and R.Rocci (eds.): *Book of Short Papers of VII Cladag Meeting*, CLEUP, 59–62.

D'Elia A. (2003a) A mixture model with covariates for ranks data: some inferential developments, *Quaderni di Statistica*, 5, 1–25.

D'Elia A. (2004) Finite sample performance of the E-M algorithm for ranks data modelling, *Statistica*, LXIII, 41–51.

D'Elia A. (2008) A statistical modelling approach for the analysis of TMD chronic pain data, *Statistical Methods in Medical Research*, 17, 389–403.

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

D'Elia A., Piccolo D. (2005b) Uno studio sulla percezione delle emergenze metropolitane: un approccio modellistico, *Quaderni di Statistica*, 7, 121–161.

D'Elia A., Piccolo D. (2006) Analyzing evaluation data: modelling and testing for homogeneity, in: S.Zani, A.Cerioli, M.Riani and M.Vichi editors: *Data Analysis, Classification and the Forward Search*, Springer, Berlin, 299–307.

Iannario M. (2007a) A statistical approach for modelling Urban Audit Perception Surveys. *Quaderni di Statistica*, 9, 149–172.

Iannario M. (2007b) Dummy variables in CUB models, *STATISTICA*, LXVIII, 2, forthcoming.

Iannario M. (2008) A class of models for ordinal variables with covariates effects, *Quaderni di Statistica*, 10, 53–72.

Iannario M. (2009a) On the identifiability of a mixture model for ordinal data, submitted for publication.

Iannario M. (2009b) A comparison of preliminary estimators in a class of ordinal data models, *Statistica & Applicazioni*, VII, 1, forthcoming.

Iannario M. (2009c) Modelling *shelter* choices in ordinal data surveys, submitted for publication.

Iannario M. (2009d) Selecting feeling covariates in rating surveys, *Italian Journal of Applied Statistics*, forthcoming.

Iannario M. (2009e) Semi-parametric estimators for ordinal data models, submitted for publication.

Iannario M. (2009f) Quantificare la percezione: un modello statistico per la valutazione del fenomeno turistico in un'area protetta, *Proceedings of IV ARIPT Congress*, Napoli-Ravello, 2007, forthcoming.

Iannario M., Piccolo D. (2009a) University teaching and students' perception: models and evidences of the evaluation process, *Proceedings of DIVAGO Meeting, University of Palermo, 2008*, Springer, forthcoming.

Iannario M., Piccolo D. (2009b) Statistical modelling of subjective survival probability, submitted for publication.

Iannario M., Piccolo D. (2009c) A program in R for CUB models inference, Version 2.0, available at *http://www.dipstat.unina.it*, this paper.

Iannario M., Piccolo D. (2010a) A new statistical model for the analysis of customer satisfaction, *Quality Technology and Quantitative Management*, 7, 149–168.

Iannario M., Piccolo D. (2010b) La percezione del tempo di attesa nei sistemi di valutazione: misura e modelli statistici, Scriptaweb, University of Salerno, forthcoming.

Piccolo D. (2003a) Computational issues in the E-M algorithm for ranks model estimation with covariates, *Quaderni di Statistica*, 5, 27–48. Corrections in: *Quaderni di Statistica*, 6, 199.

Piccolo D. (2003b) 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. (2008) Modelling University students' final grades by ordinal variables, *Quaderni di Statistica*, 10, 205–226.

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

Piccolo D., Iannario M. (2008b) Qualitative and quantitative models for ordinal data analysis, *Proceedings of MTISD 2008, Methods, Models and Information Technologies for Decision Support Systems*, Università del Salento, Lecce, 140–143.

Piccolo D., Coppola A., Del Giudice T., Capitanio F, Iannario M. (2009) Purchase behavior of typical products in modern retail, *Proceedings of* 113^{rd} *EAAE Seminar on 'A resilient European food industry and food chain in a challenging world*, Chania, Crete, Greece.

Further references related to CUB *models*

Del Giudice T., D'Elia, A. (2001) Valorizzazione dell'olio extra-vergine di oliva meridionale: una proposta metodologica per l'analisi delle preferenze, *Rivista di Economia Agraria*, LVI, 571–609.

D'Elia A. (1999) A proposal for ranks statistical modelling, in: H. Friedl, A. Berghold, G. Kauermann (eds.) *Statistical Modelling*, Graz-Austria, 468–471.

D'Elia A. (2000a) Il meccanismo dei confronti appaiati nella modellistica per graduatorie: sviluppi statistici ed aspetti critici, *Quaderni di Statistica*, 2, 173–203.

D'Elia A. (2000b) A shifted Binomial model for rankings, in Statistical Modelling - XV International Workshop on Statistical Modelling, (Nunez-Anton, V. and Ferreira, E. eds.), Servicio Editorial de la Universidad del Pais Vasco, 412–416.

D'Elia A. (2001) Efficacia didattica e strutture universitarie: la valutazione mediante un approccio modellistico, in: *Processi e metodi statistici di valutazione*, SIS, ECRA, Roma, 21–24.

D'Elia A. (2003b) Modelling ranks using the Inverse Hypergeometric distribution, *Statistical Modelling: an International Journal*, 3, 65–78.

D'Elia A., Mauriello E., Sitzia, N. (2001) Uno studio sulla scelta dei colori: esperienze e metodi statistici, *Atti del Convegno Nazionale su Matematica, Formazione Scientifica e Nuove Tecnologie*, Montevarchi, 85–96.

D'Elia A., Piccolo D. (2002a) Problemi e metodi statistici nei processi di valutazione della didattica, *Atti della Giornata di Studio su "Valutazione della Didattica e dei Servizi nel Sistema Universitario"*, Università di Salerno, 105–127.

D'Elia A., Piccolo D. (2002b) Analisi statistica delle preferenze: metodi e modelli a confronto, in: B.V. Frosini, U. Magagnoli and G. Boari, eds., *Studi in onore di Angelo Zanella*, Edizioni Vita e Pensiero, Milano, 167–187.

D'Elia A., Sitzia N. (2002) Selecting between Non-Nested Models for Preference Data, *Atti della XLI Riunione Scientifica SIS*, CLEUP, Padova, 129–132.

Piccolo D. (2007) A general approach for modelling individual choices, *Quaderni di Statistica*, 9, 31–48.