Statistical Software - Statistica & Applicazioni - Vol. XII, n. 2, 2014, pp. 205-223

NONLINEAR CUB MODELS: THE R CODE

Marica Manisera*
Paola Zuccolotto*

SUMMARY

Nonlinear CUB models have been recently introduced in the literature to model ordinal data taking into account the unequal spacing among response categories. Nonlinear CUB models can be effectively used in a variety of fields, typically when human perceptions and attitudes are measured by questionnaires with questions having ordered response categories. This paper introduces the code developed in the free software environment R for Nonlinear CUB estimation, graphical representation of a variety of outputs, fit evaluation, along with data simulation according to the Nonlinear CUB data generating process.

Keywords: *Rating Data, Likert Scales, CUB Models, EM Algorithm, Gradient Approximation Procedures, Transition Probabilities, Transition Plot.*

1. INTRODUCTION

In several fields the individuals' perceptions and attitudes are often investigated by means of Likert-type scales or, more generally, questionnaires with several questions whose possible responses are ordered. The resulting data are rating data, or ordinal data, and require appropriate statistical models to be analysed. Among the methods and techniques proposed in the literature to model rating data (among others, see Agresti, 2010; Tutz, 2012), an interesting proposal is given by CUB models (Piccolo, 2003; D'Elia and Piccolo, 2005). Several papers have been published on CUB inferential issues, identifiability problems, fitting measures, computational strategies and software routines (see Iannario and Piccolo, 2012, 2014 and the references therein). CUB models have been extended in several directions (for example, Iannario, 2012a,b,c, 2014; Grilli, Iannario, Piccolo, Rampichini, 2013; Manisera and Zuccolotto, 2014b; Piccolo, 2014) and applied in different fields (for example, Iannario, Manisera, Piccolo, Zuccolotto, 2012). This paper focuses on one possible generalization of CUB models, called Nonlinear CUB (NLCUB; Manisera and Zuccolotto, 2014a), a new class of models recently proposed in order to take account of the ordered categorical nature of the rating data. In fact, CUB models imply that the response categories are equally spaced in the respondents' mind. Instead, NLCUB can address their (possible) unequal spacing, that is when respondents, in their unconscious search for the ''right'' response category, find it easier to move, for ex-

* Dipartmento di Economia e Management - Università di Brescia - c.da S. Chiara, 50 - 25122 BRESCIA (e-mail: ✉ marica.manisera@unibs.it; paola.zuccolotto@unibs.it).

ample, from rating 1 to 2 than from rating 4 to 5. This corresponds to the concept of ''nonlinearity'' introduced by NLCUB models, defined as the nonconstantness of the transition probabilities (that is the probabilities to move from one rating to the next one). Unlike CUB, NLCUB can be used to model rating data when the transition probabilities are not constant. Further research on NLCUB is in progress and encouraged by promising results obtained in simulation studies and real data analyses (Manisera and Zuccolotto, 2013, 2014a,c, 2015a,b).

The aim of this paper is to introduce a program developed in the free software environment R for NLCUB estimation, graphical representation of a variety of outputs, fit evaluation, along with data simulation according to the NLCUB data generating process. Users can take advantage of this paper as a guide to effectively apply NLCUB models in order to obtain interesting results, also from a graphical point of view.

The paper is organized as follows: in Section 2 we describe the basic features of CUB models and the new class of NLCUB models, with a special focus on NLCUB parameter estimation (Subsection 2.1). Section 3 is devoted to the R code: the main function `NLCUB` is delineated, with the help of examples (Subsection 3.1); in addition, the single function for drawing a particular graphical output of the NLCUB model, called transition plot, and the function for data simulation according to the NLCUB data generating process are briefly described (Subsections 3.1 and 3.2). In Section 4 we present the results of a real data analysis along with the code to obtain them. Section 5 concludes. The R code is summarized in the Appendix.

2. CUB AND NONLINEAR CUB MODELS

CUB models have been introduced in the literature (Piccolo, 2003; D'Elia and Piccolo, 2005) to analyse ordinal data. With CUB, rating or ranking data are modelled by a mixture of two random variables: the observed rating r $(r = 1, \cdots, m)$ is a realization of the discrete random variable R with probability distribution given by

$$Pr\{R = r|\theta\} = \pi Pr\{V(m, \xi) = r\} + (1 - \pi)Pr\{U(m) = r\} \qquad r = 1, 2, \ldots, m$$

with $\theta = (\pi, \xi)'$, $\pi \in (0, 1]$, $\xi \in [0, 1]$. For a given m, $V(m, \xi)$ is a Shifted Binomial random variable, with trial parameter m and success probability $1 - \xi$; it models the feeling component of a decision process. $U(m)$ is a discrete Uniform random variable defined over the support $\{1, \ldots, m\}$, aimed to model the *uncertainty* component. CUB models are identifiable for $m > 3$. In terms of interpretability, $1 - \xi$ is the feeling parameter and measures the agreement with the object being evaluated, while $1 - \pi$ is the uncertainty parameter and measures the intrinsic uncertainty (indecision) in choosing the ordinal response existing in any human choice.

Nonlinear CUB models (NLCUB) are a generalization of CUB introduced by Manisera and Zuccolotto (2014a), where the NLCUB formulation is derived as a special case of a general framework describing the Decision Process (DP) that unconsciously drives individuals' responses to questions with ordered response levels. In

this general model, the DP is composed of two different approaches: (1) the feeling approach, consisting of a step-by-step reasoning, called feeling path, which proceeds through T consecutive steps. At each step, an elementary judgment is given. The last rating of the feeling path results from these elementary judgments that are summarized and transformed into a Likert-scaled rating; (2) the uncertainty approach, consisting of a random response that can be given due to the indecision surrounding any human choice, which can be related to several reasons, for example the unconscious willingness to delight the interviewer or the difficulty one can find in evaluating some specific objects. In the end, the expressed rating can derive from the feeling or the uncertainty approach with given probabilities.

With NLCUB, the discrete random variable R generating the observed rating r has a probability distribution that depends on a new parameter $T(T \geq m-1)$ and is given by

$$Pr\{R = r|\theta\} = \pi \sum_{y \in l^{-1}(r)} Pr\{V(T+1,\xi) = y\} + (1-\pi) Pr\{U(m) = r\}$$

where l is a function mapping from $(1, \cdots, T+1)$ into $(1, \cdots, m)$. In detail, l is defined as

$$l(y) = \begin{cases} 1 & \text{if} \quad y \in [y_{11}, \ldots, y_{g_1 1}] \\ 2 & \text{if} \quad y \in [y_{12}, \ldots, y_{g_2 2}] \\ \vdots & \vdots \quad \vdots \\ m & \text{if} \quad y \in [y_{1m}, \ldots, y_{g_m m}] \end{cases}$$

where y_{hs} is the h-th element of $l^{-1}(s)$, and

$$(y_{11}, \ldots, y_{g_1 1}, y_{12}, \ldots, y_{g_2 2}, \ldots, y_{1m}, \ldots, y_{g_m m}) = (1, \cdots, T+1).$$

We denote with $g_s = |l^{-1}(s)|$, where $|\cdot|$ is the cardinality of a set, the number of ''latent'' values to which rating s corresponds according to l. The values $g_1, \cdots, g_m$ univocally determine the l function. We have $T = g_1 + \cdots + g_m - 1$.

The probability distribution of a NLCUB random variable can be rewritten as

$$Pr\{R = r|\theta\} = \pi \sum_{i=g_0+\ldots+g_{r-1}}^{g_0+\ldots+g_r-1} \binom{T}{i} (1-\xi)^i \xi^{T-i} + \frac{1-\pi}{m} \tag{1}$$

with $g_0 := 0$ and $T = g_1 + \ldots + g_m - 1$. When $T = m-1$ and $g_s = 1$ for all $s = 1, \cdots, m$, the proposed model coincides with the classical CUB.

The following example provides an intuitive explanation of the functioning of the feeling approach in the NLCUB models; the formal statistics can be found in Manisera and Zuccolotto (2014a). Consider a respondent asked to express a rating, on a response scale from 1 to $m = 5$, about his job satisfaction. The idea is that the elementary judgement given at each step of the feeling path can be viewed as a quick and instinctive ''Yes/No'' response to a very simple question, related to the positive

and negative sensations that disorderly come to mind during the reasoning and together are related to the individual's job satisfaction. The simple question can be ''Do I have a positive sensation about my job satisfaction? Yes or no?'' so that the sequence of elementary judgements obtained in the feeling path is a sequence of ''Yes'' and ''No'' responses. In NLCUB models, the number of steps in the feeling path is $T > m - 1$ and the last rating of the feeling path is based on the total number of ''Yes'' responses, according to a rule determined by the l function and then by the values g_s, which denote the number of positive sensations needed to move to the next rating. As an example, we can have $T = 9$ and the rule transforming the total number of ''Yes'' responses into the last rating of the feeling path is represented in Table 2, which shows that, for example, rating 2 is reached with one, two, three or four ''Yes'' responses and moving from rating 1 to rating 2 is easier than moving from rating 2 to rating 3. In this example, $g_1 = 1$, $g_2 = 4$, $g_3 = 3$, $g_4 = 1$ and $g_5 = 1$.

TABLE 1. - *DP of NLCUB models - Feeling approach ($m = 5$ and $T = 9$)*

$T = 9$ $(> m - 1)$ elementary judgments: ''Positive sensation? Yes or no?''

Number of ''Yes'' responses	0	1	2	3	4	5	6	7	8	9
Corresponding rating	1		2				3		4	5

Due to comparability issues, the feeling parameter in NLCUB is given by the expected number μ of one-rating-point increments during the feeling path. This parameter can be interestingly interpreted as the expected value of the ratings originated within the feeling approach (see Manisera and Zuccolotto, 2014a, for details about μ and Manisera and Zuccolotto, 2013 for details about the relationship between ξ and μ). On the other hand, the uncertainty parameter is given by $1 - \pi$, like in the standard CUB. In order to interpret such parameters, one has to know that $0 \leq \mu \leq m - 1$ and $0 \leq 1 - \pi < 1$.

An interesting feature of the NLCUB model is the so-called transition probabilities $\phi_t(s)$, i.e. the probability of moving to the next rating $s + 1$ at the next step $t + 1$ of the feeling path, given that at step t the rating s has been reached, which describe the state of mind of the respondents towards the response scale used in the feeling path. Transition probabilities account for the unequal spacing between response categories, in the sense that when the probability of moving, say, from rating 1 to 2 is higher than that from rating 4 to 5, then ratings 1 and 2 can be interpreted as ''closer'' than ratings 4 and 5 in the respondents' mind. For ease of interpretation, the average transition probabilities $\phi(s)$, obtained averaging over the steps, are generally used. In this context, we also defined the unconditional probability ϕ of increasing one rating point in one step of the feeling path. Transition probabilities can be transformed, by means of a proper function h, into ''perceived distances'' between

two consecutive ratings and used for constructing the so-called transition plot, a graphical representation of the spacing existing between rating categories. Any function h able to change the transition probabilities' meaning from ''perceived closeness'' into ''perceived distance'' can be ideally chosen: for example, $h = -log(\phi(s))$ or $h = 1 - \phi(s)$. A linear transition plot suggests that the ratings are perceived as equally-spaced (all the transition probabilities are equal) while a nonlinear transition plot accounts for unequally-spaced perceived ratings. More details are in Manisera and Zuccolotto (2014a).

In order to measure the degree of nonlinearity detected by a NLCUB model, we proposed the following normalized nonlinearity index, based on the standard deviation of the transition probabilities (Manisera and Zuccolotto, 2013):

$$\lambda(\xi, \mathbf{g}) = \sigma(\phi_t(s))/\max(\sigma) \tag{2}$$

where $\mathbf{g} = (g_1, \cdots, g_m)$, $\sigma(\phi_t(s))$ is the standard deviation of $\phi_t(s)$, $\forall t, s \in \Phi$ with Φ denoting the set containing all the pairs $(t, s) : \exists\ \phi_t(s)$, and

$$\max(\sigma) = \begin{cases} \sqrt{1/4} & \text{if} \quad |\Phi| \quad \text{is} \quad \text{odd} \\ \sqrt{1/4 - 1/(4 \times |\Phi|^2)} & \text{if} \quad |\Phi| \quad \text{is} \quad \text{even} \end{cases}$$

where $|\Phi|$ denotes the cardinality of the set Φ. The index $\lambda(\xi, \mathbf{g})$ is normalized in [0,1] (or [0,100] if expressed in percentage) and can be interpreted as the proportion of nonlinearity in the NLCUB model with respect to its maximum.

2.1 *The NLCUB parameter estimation*

Given a random sample of n expressed ratings $\mathbf{s} = (s_1, \ldots, s_n)$, the loglikelihood function L of a NLCUB model for fixed $\mathbf{g} = (g_1, \ldots, g_m)$ is

$$L(\pi, \xi | \mathbf{g}; \mathbf{s}) = \sum_{i=1}^{n} \log[p_{s_i}(\pi, \xi | \mathbf{g})] = \sum_{r=1}^{m} n_r \log[p_r(\pi, \xi | \mathbf{g})] \tag{3}$$

where P_r is the probability and n_r is the frequency of rating r, respectively. The parameter estimation is performed by the following two-step procedure (Manisera and Zuccolotto, 2014a).

Step 1: Fix a maximum value T_{max} for T, and maximize (3) with respect to ξ and π, for all the possible configurations of $g_1, \cdots, g_m$ such that $g_1 + \cdots + g_m \leq T_{max} + 1$. This results in one NLCUB model for each configuration of $g_1, \cdots, g_m$, along with the corresponding ML estimates of the parameters ξ and π. The maximization of the likelihood function can be performed either by numerical optimization procedure or the EM algorithm (Manisera and Zuccolotto, 2014c).

Step 2: Among all the models defined in Step 1, select the 'best one' according

to a given criterion. Let $\hat{\mathbf{g}}$ be the configuration corresponding to the 'best' model, the parameters ξ and π are finally estimated by

$$\hat{\xi}, \hat{\pi} = \arg\max_{\xi,\pi} L(\xi, \pi|\hat{\mathbf{g}}; \mathbf{s}).$$

and the values $g_1, \cdots, g_m$ in (4) are replaced by the corresponding values in $\hat{\mathbf{g}}$.

The goodness of fit of a NLCUB model can be evaluated using some fitting measures, for example connected with the likelihood function. Among several proposals, we suggest to use one of the most popular fitting measures proposed in the framework of CUB models (see, for example, Iannario, 2009), that is the dissimilarity index $Diss = 1/2 \sum_{r=1}^{m} |fr(r) - p_r(\hat{\theta})|$, where $fr(r)$ and $p_r(\hat{\theta})$ are the observed and the expected relative frequency of response r, respectively.

According to the method for the treatment of 'don't know' (dk) responses in rating scales, proposed by Manisera and Zuccolotto (2014b), when the response scale of a question includes the *dk* option, the uncertainty parameter of a NLCUB model can be adjusted to take account of the possible presence of *dk* responses (Manisera and Zuccolotto, 2015b). The starting idea is that *dk* is a valid response to all extents and contains important information about the uncertainty of the subjects. For this reason, accounting for *dk* responses should increase the estimate of uncertainty for the whole population under study, including both respondents who are able to give an answer and those who are not. This increase is directly related to the estimated proportion of subjects in the population unable to express an evaluation. In detail, the adjusted estimate of the uncertainty parameter is simply given by $\hat{\pi}_{adj} = f_{obs}\hat{\pi}_0$, where $f_{obs} = 1 - f_{dk}$ is the relative frequency of the expressed ratings (f_{dk} is the relative frequency of *dk* responses) and $\hat{\pi}_0$ is the estimate of π obtained by fitting a NLCUB model to sample data after listwise deletion of all the *dk* responses.

More details about the parameter estimation of NLCUB models, interesting insights about the behaviour of this new class of models, suggestions on the future theoretical developments and some applications are in Manisera and Zuccolotto (2013, 2014a,b,c, 2015a,b) and the references therein.

3. THE R CODE

The code we introduce in this paper is an R script containing a number of functions, ready to be used individually or in a subset. A main function, called by the `NLCUB` command (in upper cases), recalls the other functions depending on the inputs chosen by the user. The package ''MaxLik'' (Henningsen and Toomet, 2011) must be installed. In this section, firstly a convenient code is proposed to use the main function `NLCUB` (Subsection 3.1); secondly, in Subsections 3.2 and 3.3 the single functions for drawing a transition plot and for data simulation according to the NLCUB data generating process are described. The code description is completed with examples. A summary of the R code is given in the Appendix.

3.1 *The* NLCUB *main function*

First of all, after having set the work directory, in order to cause R to accept its input from the source file ''NLCUB.r'' of the NLCUB general functions, the following code must be run:

```
source("NLCUB.r")
```

Then, the function NLCUB can be run to estimate a NLCUB model and obtain some nice results and graphical representations. The input data r to the NLCUB function can be either a sample of n expressed ratings $\mathbf{r} = (r_1, \cdots, r_n)$, with $r_i \in \{1, 2, \ldots, m\}$ for a given $m \geq 3$, or its frequency table. When the frequency table or, more precisely, the vector of the m observed frequencies is provided, the data input r must be a vector of length m, with some elements equal to 0 when the corresponding observed frequencies are null. The constant m, that is the number of possible response categories, must be specified unless vector $\mathbf{g}$ is given as input.

The function NLCUB is recalled by:

```
NLCUB(r,m=number.of.ordinal.categories,freq.table=FALSE)
```

when r is a data frame (or a vector) containing the n ratings or

```
NLCUB(r,m=number.of.ordinal.categories,freq.table=TRUE)
```

when r is a data frame (or a vector) containing the m observed frequencies, respectively. The logical flag freq.table is TRUE by default. In the latter case, the code can simply be written as NLCUB(r,m).

When the input g is not provided to the main function NLCUB, as usual, both steps in the estimation procedure are run, model selection is performed in order to determine the optimal $\hat{\mathbf{g}}$ and the final ML estimates of π and ξ are given for $\mathbf{g} = \hat{\mathbf{g}}$. The choice of the optimal $\hat{\mathbf{g}}$ is now based on the loglikelihood; other selection criteria will be soon implemented. The user can specify the maximum number maxT for T (T_{max} in Step 1 of the estimation procedure) as input in the NLCUB main function. It must be $T_{max} \geq m - 1$; if not provided, the default value maxT $= 2m - 1$ is used (as suggested in Manisera and Zuccolotto, 2013, 2014a).

Estimation of the NLCUB model can also be run for fixed $\mathbf{g}$. In this situation, the input g must be provided in the form of a vector of length m fixing the value of the ''latent'' categories $g_1, \cdots, g_m$ assigned to each rating. The ML estimates of π and ξ are computed for fixed $\mathbf{g}$ and only Step 2 of the estimation procedure is run, with $\hat{\mathbf{g}}$ set equal to the $\mathbf{g}$ provided by the user. Notice that providing fixed $g_1 = \cdots = g_m = 1$ equals to set a standard CUB model and the R code is then able to also estimate a standard CUB.

The default starting values to estimate π and ξ are $(0.5, 0.5)$. One can modify them by assigning the input param0 a vector of two different values in the parameter space (for example, param0=c(0.3,0.7)). Results in Manisera and Zuccolotto

(2014c) suggest that the naive choice of starting from (0.5, 0.5) is a good strategy, which generally allows to reach the loglikelihood function global maximum, and mention the need of accurate starting values, which remains an open issue in the context of NLCUB models.

The user can also choose the maximization procedure to be used in the Step 1 of the estimation. Up to now, two main methods are implemented: (1) by giving `method="NM"` as input, the main function `NLCUB` run the Nelder-Mead numerical optimization procedure contained in the R package ''MaxLik'' (Henningsen and Toomet, 2011), which comprises several other maximization methods: Newton-Raphson, Broyden-Fletcher-Goldfarb-Shanno, i.e., the BFGS algorithm implemented in R, Berndt-Hall-Hall-Hausman, Simulated Annealing, Conjugate Gradients (see, for example, Greene, 2008 and the references therein and in Henningsen and Toomet, 2011); (2) by giving `method="EM"` as input, the EM algorithm is used, according to the suggestions in the literature on finite mixture models. The default method is `method="EM"`.

In the end, the numerical outputs of the main function `NLCUB` when the maximization procedure is the Nelder-Mead numerical optimization (`method="NM"`) are:

- parameter estimates `pai` for π and `csi` for ξ
- the optimal `g` for $\mathbf{g} = [g_1, \cdots, g_m]$ (if $\mathbf{g}$ is not declared as input);
- the estimated asymptotic variance-covariance matrix `Varmat` of the ML estimator for (π, ξ) for fixed $\mathbf{g}$;
- the estimated Information matrix `Infmat`;
- the m fitted frequencies `Fit`, obtained according to the NLCUB model with parameters `pai`, `csi` and `g`;
- the dissimilarity index `diss`;
- the transition probability matrix `transprob_mat`, giving the estimates of the transition probabilities $\phi_t(s)$, that is the probability to move from one rating to the next one in the feeling path, computed at each step t of the feeling path;
- the $m - 1$ transition probabilities `transprob`, that is the estimates of the probabilities $\phi(s)$ to move from ratings s, $s = 1, \ldots, m-1$ to rating $s + 1$, computed averaging over the steps the probabilities in the transition probability matrix;
- the unconditioned transition probability `uncondtransprob`, that is the estimate of the probability ϕ of increasing one rating point in one step in the feeling path, independently on the rating reached at the previous step;
- the expected number of one-rating-point increments during the feeling path `mu`, estimate of μ, that is interpreted as the feeling parameter in a NLCUB model;
- the nonlinearity index `NL_index` (λ);
- the estimate `pai_adj` of the uncertainty parameter adjusted for the presence of ''don't know'' (*dk*) responses (Manisera and Zuccolotto, 2015b). In order to get this result, the proportion of *dk* responses observed in the data must be given in the `dk` input.

EXAMPLE 1

Consider a number $n = 500$ of customers asked to rate their satisfaction with a certain product on a response scale ranging from 1: ''very dissatisfied'' to 5: ''very satis-

fied''. The resulting sample of n ratings can be stored in a n-dimensional vector or, alternatively, in a m-dimensional vector of the frequencies corresponding to the m response categories. Obviously, data can be available in an external file that can be read in R by a proper function. In this example, the vector of the $m = 5$ observed frequencies is $(52, 79, 149, 131, 89)$. The code to estimate the NLCUB model with Nelder-Mead numerical optimization procedure (`method="NM"`) is the following.

```
r <- c(52,79,149,131,89)

number.of.ordinal.categories <- 5

NLCUB(r,m=number.of.ordinal.categories,freq.table=TRUE,
method="NM")
```

The output is given below.

```
[[1]]
Maximum Likelihood estimation
Nelder-Mead maximisation, 59 iterations
Return code 0: successful convergence
Log-Likelihood: -773.6971 (2 free parameter(s))
Estimate(s): 0.4311235 0.1876388

$pai [1] 0.4311235

$csi [1] 0.1876388

$Varmat

              [,1]           [,2]
[1,] 2.767150e-03  -1.631647e-05
[2,]-1.631647e-05   2.459070e-04

$Infmat

              [,1]           [,2]
[1,]    0.72304829     0.04797585
[2,]    0.04797585     8.13633960

$g [1] 1 5 2 1 1

$Fit

           [,1]
[1,]  0.1137754
```

```
[2,]  0.1440378
[3,]  0.3101198
[4,]  0.2518647
[5,]  0.1802022

$diss [1] 0.02409751

$transprob [1] 0.8123612 0.2362724 0.2532141 0.8123612

$transprob_mat

          [,1]       [,2]       [,3]       [,4]
[1,] 0.8123612        NaN        NaN        NaN
[2,] 0.8123612  0.0000000        NaN        NaN
[3,] 0.8123612  0.0000000        NaN        NaN
[4,] 0.8123612  0.0000000        NaN        NaN
[5,] 0.8123612  0.0000000        NaN        NaN
[6,] 0.8123612  0.2874723        NaN        NaN
[7,] 0.8123612  0.4541016  0.0000000        NaN
[8,] 0.8123612  0.5462511  0.3104340        NaN
[9,] 0.8123612  0.6023540  0.4492084  0.8123612

$uncondtransprob [1] 0.2842515

$mu [1] 2.558263

$NL_index [1] 0.648814

$pai_adjusted_for_dk [1] Parameter pai has not been adjusted for dk
responses
```

Results show that the ML estimates of π and ξ are 0.43 and 0.19, respectively. This means that the customers rated their satisfaction with a moderately high uncertainty ($1 - \pi = 0.57$). The feeling component in NLCUB model is measured by μ, whose estimate results 2.56. This means that during the feeling path the expected number of one-rating-point increments equals 2.56. Since $0 \leq \mu \leq 3$, the feeling can be considered fairly high and, then, customers are highly satisfied with the product under evaluation. The estimated values of $\mathbf{g}$ suggest that moving from rating 1 to 2 is easier than moving from 2 to 3 as well as from 3 to 4, while moving from 4 to 5 returns to be easy. This is confirmed by the estimated average transition probabilities (`transprob`), which are a measure of the perceived closeness between two adjacent response categories. The matrix of the estimated transition probabilities (`transprob_mat`) also confirms this result. It is a $T \times (m-1)$ matrix and the generic element in row t, $t = 1, \cdots, m$, and column s,

$s = 1, \cdots, m-1$, is given by the estimate of $\phi_t(s)$: for example, the estimated $\phi_6(2)$ is 0.29 and this means that the probability of moving to rating 3 at the step 7 of the feeling path, given that at step 6 the rating 2 has been reached, is equal to 0.29. Some of the elements of this matrix are not defined (`NaN`): for example, at step 1 of the feeling path, it is not possible to reach ratings 3, 4, or 5.

When the EM algorithm is used to estimate a NLCUB model (`method="EM"`, the default), the numerical outputs of the main function `NLCUB` listed above are enriched by a list of several other results:

- the lists of the parameter estimates `pailist` for π and `csilist` for ξ obtained at each step of the EM algorithm;
- the maximum value `maximum` of the log-likelihood function obtained at the end of the EM iterations;
- the number `k` of iterations used in the EM algorithm;
- a return code `check` to assess whether the EM algorithm stops for a successful convergence (`check=successful convergence`) or because the maximum number of iterations is reached (`check=iterations stopped (maxiter)`).

When the EM algorithm is used, the analytic formulas to derive the estimates of the asymptotic variance-covariance matrix and the Information matrix (Manisera and Zuccolotto, 2014c) are needed. On the contrary, when the numerical algorithm for optimization implemented in the `method='NM'` is used, the variance-covariance matrix for the current estimates can be immediately derived as usual: in this case, the R code provides the estimated asymptotic variance-covariance matrix `Varmat` and the estimated Information matrix `Infmat` derived from the value approximating the Hessian at the parameter values where convergence has occurred.

EXAMPLE 2

Considering the data in Example 1, the code to estimate the NLCUB model with EM algorithm (the default `method`) `is the following`.

```
r <- c(52,79,149,131,89)

number.of.ordinal.categories <- 5
NLCUB(r,m=number.of.ordinal.categories,freq.table=TRUE)
```

Among the interesting outputs of the main function `NLCUB`, there are two plots, the observed vs. fitted frequencies and the transition plot. If one prefer not to display graphs, the logical flag `draw.plot` in input to the main function `NLCUB` must be set to `FALSE` (default is `TRUE`). Figure 1 shows the plots obtained with the code given in the Example 1. The left panel plot shows that the fit of the expected to the observed relative frequencies is very good ($Diss = 0.0218$); the obtained nonlinear transition plot (right panel) shows that respondents find it easier moving from rating 1 to 2, for example, than from rating 2 to 3.

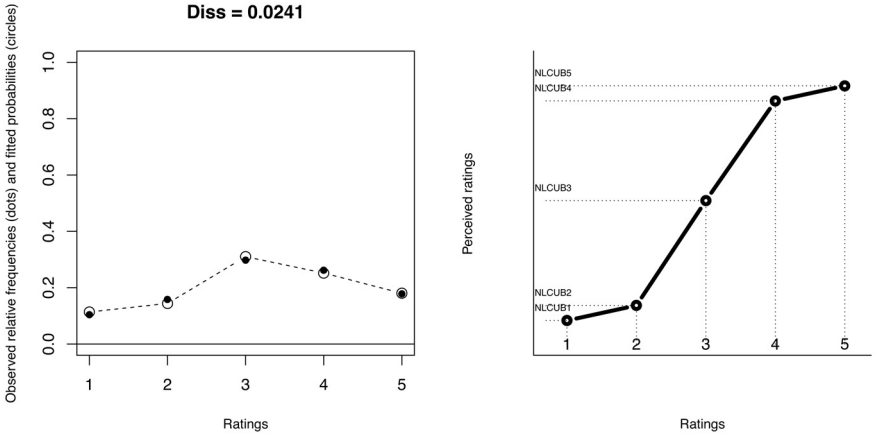

FIGURE 1. - *Example 1 - Observed vs. fitted frequencies plot (left) and the transition plot (right)*

3.2 *The* `transplot` *function for drawing a transition plot*

A transition plot of a NLCUB model can be easily obtained with the following code:

```
transplot(xip2,gp2,log.scale=TRUE)
```

where `xip2` is the value of the ξ parameter and `gp2` is the vector $\mathbf{g}$ of the ''latent'' categories assigned to each rating point. These two values can derive from an estimated NLCUB or can be fixed by the user. The `log.scale` logical flag refers to the scale used to transform the transition probabilities into perceived distances to be plotted: when `TRUE` (default), the logarithmic transformation is used ($h = -log(\phi(s))$) while when it is `FALSE`, the linear scale is used ($h = 1 - \phi(s)$).

3.3 *The* `simul` *function for data simulation*

In order to simulate *Nsim* pseudo-random numbers from a NLCUB model with given ξ, π, $\mathbf{g}$, the following code can be run:

```
simul(Nsim,paisim,xisim,gsim)
```

where `Nsim` is the sample size and `paisim, xisim` and `gsim` are, respectively, the values of parameters π, ξ, $\mathbf{g}$ used for simulation. The output of this function is the vector of *Nsim* ratings, simulating according to a NLCUB model with the chosen values of the parameters. This function turns out to be useful especially for further research developments on NLCUB models.

4. A REAL DATA ANALYSIS

In this section, we analyse real data from the Standard Eurobarometer 78, a sample survey covering the national population of citizens of the 27 European Union member states. The number of interviewees ranges from 500 (Malta) to 1,561 (Germany), with an average of 986 over the 27 countries. More details on the Standard Eurobarometer 78 can be found on the European Union website (Eurobarometer 78.1 (2012): TNS Opinion & Social, Brussels, available from http://ec.europa.eu/public_opinion/archives/eb/eb78/eb78_en.htm).

The same application was published in Manisera and Zuccolotto (2014a), where a comparison between the results from NLCUB models and traditional CUB was also presented.

The focus in on one single question of the Eurobarometer questionnaire (QA3.2: ''How would you judge the current situation of the European economy?'') and three selected countries (Greece, Germany, Italy). The ratings were expressed on a Likert scale with $m = 4$ possible responses (''very bad'', ''rather bad'', ''rather good'', ''very good''). The ''don't know'' option has been treated with listwise deletion; a procedure to adjust the estimate of the uncertainty in the model in order to account for the ''don't know' responses has been proposed in Manisera and Zuccolotto (2014b, 2015b) and implemented for both CUB and NLCUB models in the R code for NLCUB models.

We fitted data with NLCUB model setting $T_{max} = 2m - 1 = 7$. The code to be run to obtain both numerical and graphical results for Greece, for example, is the following:

```
source("NLCUB.r")
r <- c(390,440,150,10)
number.of.ordinal.categories <- 4
NLCUB(r,m=number.of.ordinal.categories,freq.table=TRUE,method="NM")
```

Figures 2-4 show the observed relative frequencies and the corresponding NLCUB fitted probabilities (left panels) and the transition plots (right panels) for Greece, Germany and Italy, respectively. The fit is quite good, as confirmed also by the values of the *Diss* index, reported on top of the left panels in Figures 2-4.

The transition plots implied by the estimated NLCUB models for the three countries reveal that Greece shows a linear transition plot, whilst Germany and Italy show two slightly different nonlinear DPs, both characterized by a decreasing probability of moving to higher ratings. This means that, in general, for Greek respondents moving, for instance, from rating 1 to 2 is as hard as moving from rating 3 to 4. German and Italian respondents find it easier moving from rating 1 to 2 than moving from rating 3 to 4.

The NLCUB parameter estimates are reported in Table 5. As expected, for Greece the estimated NLCUB exactly matches the CUB structure ($g_s = 1$, $s = 1, \cdots, 4$) and $\lambda = 0$. On the other hand, for Germany and Italy, the obtained values of $g_1, \cdots, g_4$

and λ confirm the nonlinear structure shown in the transition plots; as expected, the transition probabilities decrease when moving towards the higher ratings.

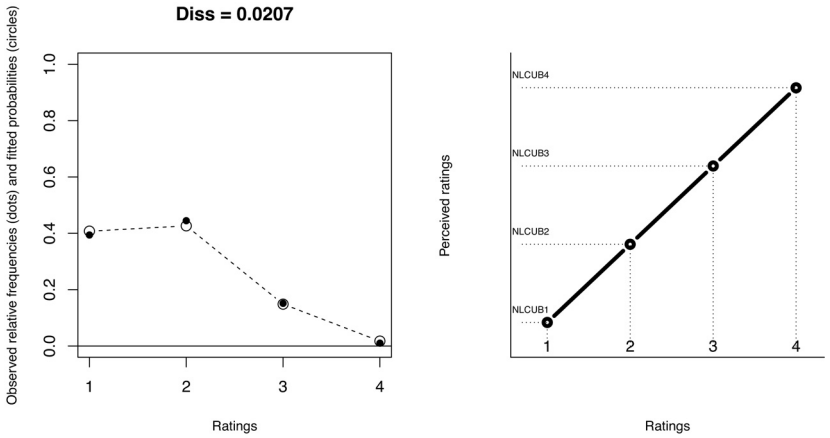

FIGURE 2. - *Greece: Observed relative frequencies vs NLCUB fitted probabilities (left); transition plot (right), Standard Eurobarometer 78 (QA3.2)*

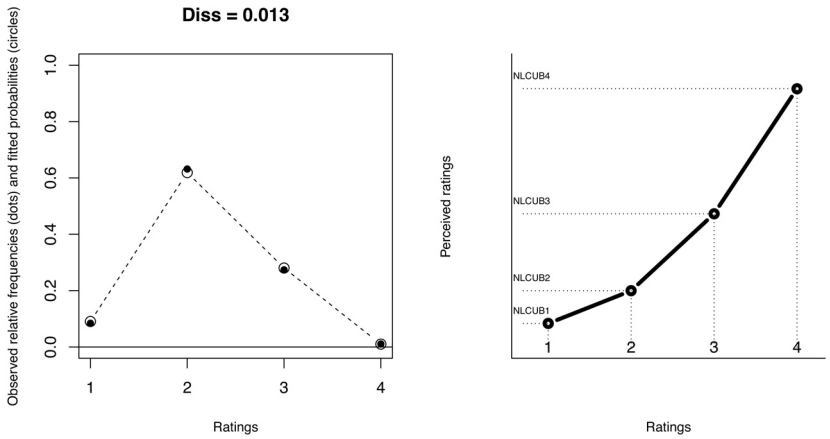

FIGURE 3. - *Germany: Observed relative frequencies vs NLCUB fitted probabilities (left); transition plot (right), Standard Eurobarometer 78 (QA3.2)*

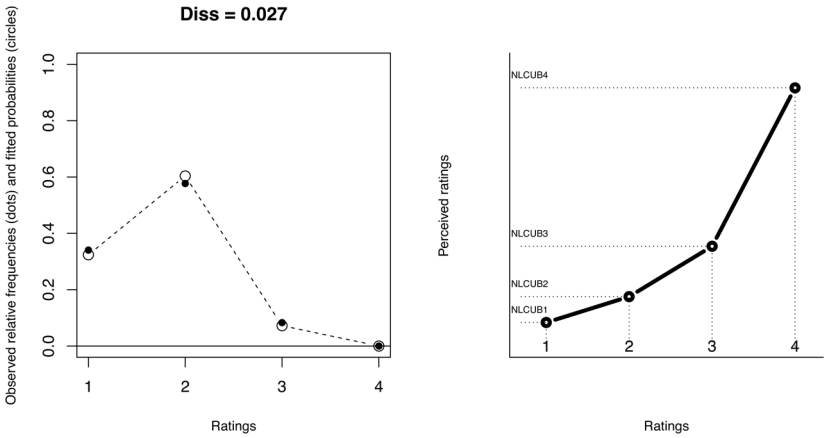

FIGURE 4. - *Italy: Observed relative frequencies vs NLCUB fitted probabilities (left); transition plot (right), Standard Eurobarometer 78 (QA3.2)*

On the whole, all respondents from the three countries exhibit a very low uncertainty. Some differences can be observed in the results concerning the feeling: the estimated μ suggest that Italian respondents are the most pessimistic about the European economy, immediately followed by Greek respondents, while German respondents are more confident.

TABLE 2. - *NLCUB parameter estimates, Standard Eurobarometer 78 (QA3.2), Greece, Germany and Italy*

Greece

ξ	π	g_1	g_2	g_3	g_4
0.7413	0.9999	1	1	1	1
$\phi(1)$	$\phi(2)$	$\phi(3)$	ϕ	μ	λ
0.2587	0.2587	0.2587	0.2587	0.7761	0.0000

Germany

ξ	π	g_1	g_2	g_3	g_4
0.6167	0.9925	1	2	2	1
$\phi(1)$	$\phi(2)$	$\phi(3)$	ϕ	μ	λ
0.3833	0.1057	0.0258	0.2416	1.2080	0.3064

Italy

ξ	π	g_1	g_2	g_3	g_4
0.8512	0.9999	1	2	4	1
$\phi(1)$	$\phi(2)$	$\phi(3)$	ϕ	μ	λ
0.1487	0.0249	8.7e-06	0.1069	0.7484	0.1306

5. CONCLUSIONS AND FUTURE DEVELOPMENTS

In this paper, after a brief presentation of the NLCUB models, we have introduced the main features of the R code useful to estimate the NLCUB models, obtain some related graphical representations and simulate data with a NLCUB data generating process.

For what concerns computational time, one model is estimated in tenths of a second. The convergence rate of the EM algorithm is about 10 times that of NM. This could be accelerated by including in the procedure the use of preliminary estimators for the parameters of the model, as proposed for standard CUB by Iannario (2012d).

The R code is freely available at https://www.researchgate.net/publication/277717439, so that NLCUB models can be easily applied in several fields, whenever the aim is to analyze rating data coming, from example, from the administration of questionnaires with ordinal response scales, also known as Likert-type scales. The available R code is extremely simple to use and does not require strong technical skills. Despite that, the proposed R program allows any interested user to enter all the single functions composing the main function, which can be exploited and customized in order to analyze in further detail the features of the NLCUB models and implement methodological variations in the models.

Our future research will be devoted to make the R code suitable to be released as a standard R package. Since the research about NLCUB starts from the study of the standard CUB and is strongly related to CUB models and their extensions, our R code could become part of a very general R package on CUB and its generalizations.

ACKNOWLEDGEMENT

The authors thank Prof. Domenico Piccolo and Prof. Maria Iannario for many stimulating discussions. This research was partially funded by STAR project (University of Naples Federico II - CUP: E68C13000020003) and partially by a grant from the European Union Seventh Framework Programme 'Cooperation - Socio-Economic Sciences and Humanities' (FP7-SSH/2007-2013); *'SYstemic Risk TOmography: Signals, Measurements, Transmission Channels, and Policy Interventions' - SYRTO* - Project ID: 320270.

APPENDIX

In this Appendix, the R code introduced in this article is shortly reported in the form of help pages in R.

DESCRIPTION

Generic code for Nonlinear CUB estimation, graphical representations, fit evaluation, data simulation

USAGE

```
NLCUB(r,g = c(), m = c(), maxT = c(), param0 = c(0.5,0.5), freq.table
= TRUE, method = "EM", draw.plot = TRUE, dk = c() )
```

ARGUMENTS

`r`	a vector of observed ratings (either microdata or the m observed frequencies - frequency table); see `freq.table`
`m`	integer: number of categories of the response scale (active only when `g` is not declared)
`g`	a vector of the ''latent'' categories assigned to each rating point; if `g` is declared, Nonlinear CUB parameters are estimated for fixed **g**, else model selection is performed in order to determine the optimal **g**
`maxT`	integer: maximum value for T (must be `maxT` $> m - 1$, default is $2m - 1$) (active only when `g` is not declared)
`param0`	starting values for π and ξ
`freq.table`	logical: if `TRUE`, the data in `r` is the vector of the m observed frequencies (frequency table)
`method`	character: method to use for likelihood maximization; `method="NM"` for likelihood based - Melder-Mead maximization - `method="EM"` for likelihood based - EM algorithm
`draw.plot`	logical: if `TRUE`, two graphs are plotted: observed vs fitted frequencies and transition plot
`dk`	proportion of ''don't know'' responses; if declared, in addition to the estimate of π, the estimated of π adjusted for the presence of *dk* responses is provided

VALUE

`pai`	parameter estimate for π
`csi`	parameter estimate for ξ
`g`	optimal value for $\mathbf{g} = [g_1, \cdots, g_m]$ (if `g` is not declared as input)
`Varmat`	estimated asymptotic variance-covariance matrix of the ML estimator for (π, ξ) for fixed **g**
`Infmat`	estimated Information matrix

`Fit`	m fitted frequencies, obtained according to the estimated NLCUB model
`diss`	the dissimilarity index value
`transprob_mat`	transition probability matrix containing the estimates of $\phi_t(s)$
`transprob`	$m-1$ estimated transition probabilities $\phi(s)$
`uncondtransprob`	estimate of unconditioned transition ϕ probability
`mu`	estimate of μ
`NL_index`	the nonlinearity index value
`pai_adj`	estimate of the uncertainty parameter adjusted for the presence of ''don't know'' (dk) responses
`pailist`	list of the parameter estimates for π obtained at each step of the EM algorithm (active only when `method="EM"`)
`csilist`	list of the parameter estimates for ξ obtained at each step of the EM algorithm (active only when `method="EM"`)
`maximum`	the maximum value of the log-likelihood function obtained at the end of the EM iterations (active only when `method="EM"`)
`k`	the number of iterations used in the EM algorithm (active only when `method="EM"`)
`check`	a return code to assess whether the EM algorithm stops for a successful convergence or because the maximum number of iterations is reached (active only when `method="EM"`)

REFERENCES

Agresti A. (2010). *Analysis of ordinal categorical data, 2nd ed.*. Wiley, Hoboken.

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

Greene W. (2008). *Econometric analysis, 6th ed.*. Prentice Hall, New York.

Grilli L., Iannario M., Piccolo D., Rampichini C. (2013). Latent class CUB models. *Advances in Data Analysis and Classification*, **8**, 105-119.

Henningsen A., Toomet O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, **26**, 443-458.

Iannario M. (2009). Fitting measures for ordinal data models. *Quaderni di Statistica*, **11**, 39-72.

Iannario M. (2012a). Hierarchical CUB models for ordinal variables. *Communications Statistics - Theory and Methods*, **41**, 3110-3125.

Iannario M. (2012b). Modelling shelter choices in a class of mixture models for ordinal responses. *Statistical Methods & Applications*, **20**, 1-22.

Iannario M. (2012c). CUBE models for interpreting ordered categorical data with overdispersion. *Quaderni di Statistica*, **14**, 137-140.

Iannario M. (2012d). Preliminary estimators for a mixture model of ordinal data. *Advances in Data Analysis and Classification*, **6**, 163-184.

Iannario M. (2014). Modelling uncertainty and overdispersion in ordinal data. *Communications in Statistics - Theory and Methods*, **43**, 771-786.

Iannario M., Manisera M., Piccolo D., Zuccolotto P. (2012). Sensory analysis in the food industry as a tool for marketing decisions. *Advances in Data Analysis and Classification*, **6**, 303-321.

Iannario M., Piccolo D. (2012). CUB models: statistical methods and empirical evidence. In R.S. Kenett, S. Salini (Eds.), *Modern Analysis of Customer Surveys* (pp. 231-258). Wiley, New York.

Iannario M., Piccolo D. (2014). *Inference for CUB models: a program in R. Statistica & Applicazioni*, **XII**, 2, 177-204.

Manisera M., Zuccolotto P. (2013). Nonlinear CUB models: some stylized facts. *QdS - Journal of Methodological and Appllied Statistics*, **1(2)**.

Manisera M., Zuccolotto P. (2014a). Modeling rating data with Nonlinear CUB models. *Computational Statistics & Data Analysis*, **78**, 100-118.

Manisera M., Zuccolotto P. (2014b). Modeling 'don't know' responses in rating scales. *Pattern Recognition Letters*, **45**, 226-234.

Manisera M., Zuccolotto P. (2014d). Numerical optimization and EM algorithm in a mixture model for human perceptions analysis. *Working paper*.

Manisera M., Zuccolotto P. (2015a). Identifiability of a model for discrete frequency distributions with a multidimensional parameter space. *Journal of Multivariate Analysis*, **140**, 302-316.

Manisera M., Zuccolotto P. (2015b). Treatment of 'don't know' responses in a mixture model for rating data. *Metron*, forthcoming.

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

Piccolo D. (2014). Inferential issues on CUBE models with covariates. *Communications in Statistics - Theory and Methods*, **43**, forthcoming.

Tutz G. (2012). *Regression for categorical data.* Cambridge University Press, Cambridge.