QdS – Vol. 15, 2013

Nonlinear CUB models: some stylized facts

Marica Manisera
University of Brescia
E-mail: manisera@eco.unibs.it

Paola Zuccolotto
University of Brescia
E-mail: zuk@eco.unibs.it

Summary: The Nonlinear CUB models have been recently introduced with the aim of generalizing the standard CUB in the context of rating data modelling. In this paper the stylized facts concerning the main features of the Nonlinear CUB models are established by means of an extended systematic analysis of a great number of different models. Results provide interesting insights on this new class of models and suggestions about the future theoretical developments.

Keywords: CUB models; Nonlinear CUB models; Rating data; Likert-type scales; Latent variables; Transition probability; Transition plot.

1. Introduction

Statistical analyses in several fields often deal with rating data, used to investigate the individuals' perceptions, attitudes, behaviours, cognitions. Rating data are usually collected by means of a questionnaire involving categorical ordinal items, i.e. questions whose possible responses are measured on an ordinal scale. In the literature, several methods and techniques have been proposed to model rating data, taking into account their categorical ordinal nature (see Agresti, 2010; Tutz, 2012). Among them, a different paradigm is given by the CUB models (D'Elia and Piccolo, 2005; Piccolo, 2006; Piccolo and D'Elia, 2008; Iannario and Piccolo, 2012), introduced in 2003 with the name MUB (Piccolo, 2003). Since then, the CUB models have been developed in several directions and many papers concerning inferential issues, identifiability problems, fitting measures, computational strategies and software routines have been published (Iannario, 2009, 2010, Iannario and Piccolo, 2010, 2012, 2014). In addition, the CUB models have been extended in several directions, for example to consider subjects' and objects'

covariates (Iannario 2007, 2008; Piccolo, 2013), the so-called shelter effect, resulting in a very high frequency on a given response category (Iannario, 2012a), the possible presence of a hierarchical structure in the data (Iannario, 2012b), multimodal response distributions deriving from a latent class structure (Grilli, Iannario, Piccolo and Rampichini, 2013). The interest towards the CUB models has increased also from the point of view of applications, because they can be used effectively in different contexts, for example linguistics (Balirano and Corduas, 2008), risk analysis (Cerchiello, Iannario and Piccolo, 2010), marketing (Iannario, Manisera, Piccolo e Zuccolotto, 2012), medicine (D'Elia, 2008), sensometrics (Piccolo and D'Elia, 2008).

A possible generalization of the CUB models is the so-called Nonlinear CUB (NLCUB), a new class of models recently proposed in order to deal with the unequal spacing of the response categories in the respondent's mind (Manisera and Zuccolotto, 2014). The unequal spacing of categories has been translated into the concept of nonlinearity, defined as the presence of non-constant transition probabilities, i.e. the probabilities of moving from one rating to the next one during a decision process where the expressed rating derives from a step-by-step mechanism. NLCUB, differently from CUB, can be used to model rating data with non-constant transition probabilities. Simulation studies and real data analyses (Manisera and Zuccolotto, 2013, 2014) show promising results that encourage further research.

The aim of this paper is to present some stylized facts concerning the NLCUB models, deriving from an extended systematic study performed in order to investigate their behaviour. Based on the findings of this study, we draw some interesting conclusions on the nonlinearity patterns expressed by different NLCUB models and useful suggestions concerning their estimation procedure.

The paper is organized as follows: in Section 2 we describe the basic features of the CUB and NLCUB models. In particular, the concept of transition probability is defined in Subsection 2.2 and its formulation is derived for both CUB and NLCUB models, while the parameter estimation procedure of NLCUB models is briefly recalled in Subsection 2.3. In Section 3 we introduce the concepts of linearity and nonlinearity of the decision process underlying the responses on a rating scale. Also, we propose a nonlinearity index able to measure the degree of nonlinearity in the decision process. In Section 4 the results of a wide systematic study are presented and summarized by some stylized facts. Section 5 concludes the paper.

2. CUB and Nonlinear CUB models

The CUB models have been introduced in the literature to analyse ordinal data and fit in the latent variable framework. With the CUB models, rating or ranking data are modelled by a mixture of a Uniform and a Shifted Binomial random variables: the observed rating r ($r = 1, \ldots, m$) is a realization of the discrete random variable R

whose probability distribution is given by

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

with $r = 1,\ldots,m$, $\theta = (\pi,\xi)'$, $\pi \in (0,1]$, $\xi \in [0,1]$. The number of possible response categories m is a given and known integer. For a given m, $V(m,\xi)$ is a Shifted Binomial random variable, with trial parameter m and success probability $1-\xi$, modelling the *feeling* component, and $U(m)$ is a discrete Uniform random variable defined over the support $\{1,\ldots,m\}$, aimed to model the *uncertainty* component. The CUB models are identifiable for $m > 3$ (Iannario, 2010).

The Nonlinear CUB models (NLCUB), introduced by Manisera and Zuccolotto (2014), are a generalization of the CUB models. In detail, with NLCUB the discrete random variable *R* generating the observed rating r has a probability distribution depending on a new parameter T, $T \geq m-1$ and given by

$$Pr\{R = r|\theta\} = \pi \sum_{y \in l^{-1}(r)} Pr\{V(T+1,\xi) = y\} + (1-\pi)P\{U(m) = r\} \tag{2}$$

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

$$l(y) = \begin{cases} 1 & \text{if} \quad y \in [y_{11},\ldots,y_{g_1 1}] \\ 2 & \text{if} \quad y \in [y_{12},\ldots,y_{g_2 2}] \\ \vdots & \vdots \quad \vdots \\ m & \text{if} \quad y \in [y_{1m},\ldots,y_{g_m m}] \end{cases} \tag{3}$$

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

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

We denote with $g_s = |l^{-1}(s)|$, where $|\cdot|$ denotes the cardinality of a set, the number of "latent" values to which rating s corresponds based on l. The values $g_1,\ldots,g_m$ univocally determine the function l and can be considered as parameters of the model. We have $T = g_1 + \ldots + g_m - 1$.

When $T = m-1$ and $g_s = 1$ for all $s = 1,\ldots,m$, then the proposed model collapses into the standard CUB model.

2.1. The general framework for the decision process

In Manisera and Zuccolotto (2014) the NLCUB formulation is derived as a special case of a more general framework, proposed to describe the decision process (DP) driving individuals' responses to survey questions with ordered response levels. This general model assumes the presence of two different approaches, which compose the DP and,

borrowing the CUB terminology, are called feeling and uncertainty approach, respectively. The feeling approach proceeds through T consecutive steps, called feeling path. At each step, an elementary judgment is given. The rating of the feeling path results from these elementary judgments that are, firstly, summarized and, secondly, transformed into a Likert-scaled rating. The uncertainty approach consists of a random judgment that can be given by the respondent due to the indecision in choosing the ordinal response, depending on a great variety of possible reasons, e.g. unconscious willingness to delight the interviewer, difficulty in evaluating some specific objects using limited information, partial understanding, lack of self-confidence, laziness, boredom, etc. In the end, the expressed rating can derive from the feeling or the uncertainty approach with given probabilities. Some existing statistical models can be viewed as special cases of this general framework.

The most interesting feature of this DP is the mechanism that, along the feeling path, generates the rating according to the feeling approach. We address the reader to the seminal paper on NLCUB models for a formal statistical description and two illustrative examples that highlight the difference between the CUB and NLCUB models. Here we limit ourselves to provide an intuitive explanation. First of all, the difference between the DPs of NLCUB and CUB models only pertain the feeling approach. In both models, 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. For example, when a respondent is asked to express his/her agreement with a certain statement by using a Likert scale from 1 to $m = 5$, the simple question can be "Do I have a positive sensation about this statement? Yes or no?". The sequence of elementary judgements obtained in the feeling path is a sequence of "Yes" and "No" responses that reflect the set of positive and negative sensations that disorderly come to mind during the reasoning, according to the individual's experience about the latent trait being evaluated. The main difference between CUB and NLCUB is that:

1. in the DP of CUB models, the number of steps in the feeling path (that is the number of simple questions) is $T = m - 1 = 4$ and the last rating of the feeling path is given by 1 plus the total number of "Yes" responses. This last rating follows a Shifted Binomial distribution, since the basic judgments are realizations of *iid* Bernoulli random variables;

2. in the DP of NLCUB models, the number of steps in the feeling path is $T > m-1$ and the last rating of the feeling path is still based on the total number of "Yes" responses, but in an unbalanced way. As an example, we can have $T = 9$ and the total number of "Yes" responses can be transformed into the last rating of the feeling path by the rule represented in Table 1, 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 much more easier than moving from rating 2 to rating 3.

Then, in both CUB and NLCUB the final response can be either the rating deriving from the feeling approach or a random rating resulting from the uncertainty approach,

Table 1. DP of NLCUB models - Feeling approach (example with $m = 5$ 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

with probabilities π and $1 - \pi$, respectively. It is easy to see that the expressed ratings derived from the mechanism in 1. and in 2. follow distribution (1) and (2), respectively. In particular, in (2) the asymmetric correspondence between the total number of "Yes" responses and the rating of the feeling approach is accounted for by the function l and the values g_s, which denote the number of positive sensations needed to move to the next rating (in the above example, $g_1 = 1$, $g_2 = 4$, $g_3 = 3$, $g_4 = 1$ and $g_5 = 1$).

2.2. Transition probabilities for CUB and Nonlinear CUB

The way respondents achieve, moving through T steps, the formulation of a rating in the feeling approach is called feeling path. The examples of DP considered in the previous section show that, in the end of the feeling path, the respondent (unconsciously) considers the total number of "Yes" responses (i.e. the total number of positive sensations that came into his/her mind) and decides which rating should be assigned, according to some rule. As a matter of fact, we can imagine that the same reasoning is made at each step of the feeling path. In other words, at each step (i.e. for each new basic judgment he/she expresses), the respondent considers the number of "Yes" responses collected up to that moment and formulates a provisional rating, which will be updated at the next step, until the T-th step has been reached.

Within this framework, we can express the so-called transition probabilities $\phi_t(s)$, i.e. the probability of moving to provisional rating $s + 1$ at step $t + 1$ of the feeling path, given that the provisional rating at step t is s, $s = 1, \ldots, m - 1$. Transition probabilities depend on the function l, i.e. on the rule according to which the respondents transform the number of "Yes" responses into the rating during the feeling path. Therefore, transition probabilities describe the respondents' state of mind about the response scale used to express judgments in the feeling approach.

For ease of interpretation, the average transition probability $\phi(s)$, obtained averaging $\phi_t(s)$ over t, is generally used. It indicates the "perceived closeness" between ratings s and $s+1$ and can be transformed into a "perceived distance" $\delta_s = h(\phi(s))$ by means of a proper function (usually $\delta_s = -\log(\phi(s))$). These quantities are the basis for constructing the so-called transition plot, useful to detect whether the ratings are perceived by respondents as equally spaced or not. In the transition plot, a broken line joins the points $(s, \tilde{\phi}(s-1))$, $s = 1, \ldots, m$, $\tilde{\phi}(0) = 0$, and $\tilde{\phi}(s-1) = (\delta_1 + \cdots + \delta_{s-1})/(\delta_1 + \cdots + \delta_{m-1})$ for $s = 2, \ldots, m$. Figure 1 represents two examples of linear (left) and nonlinear (right)

transition plot.

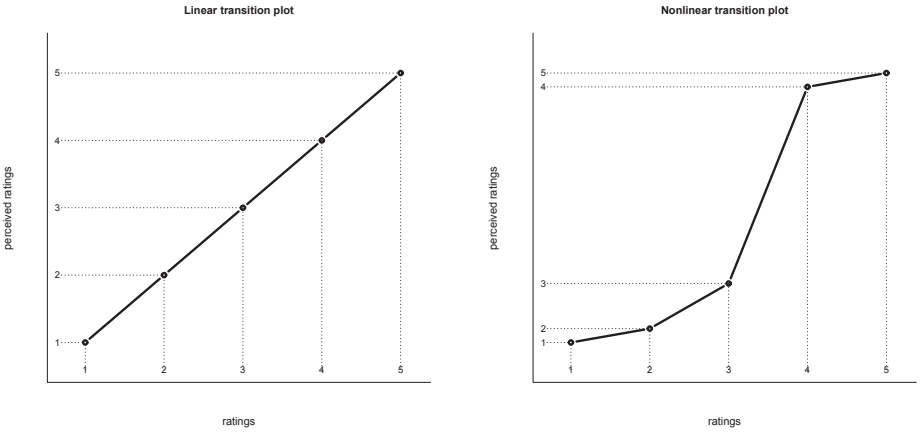

Figure 1. Examples of linear (left) and nonlinear (right) transition plot ($m = 5$)

By construction, the y-axis in the transition plot ranges in $[0, 1]$. A linear transition plot suggests that the ratings are perceived as equally-spaced in the respondents' mind (Figure 1, left) while a nonlinear transition plot accounts for unequally-spaced perceived ratings (Figure 1, right).

Starting from the transition probabilities, we can also define the expected number μ of one-rating-point increments during the feeling path and the unconditional probability of increasing one rating point in one step of the feeling path $\phi = \mu/T$.

Manisera and Zuccolotto (2014) derive $\phi_t(s)$, $\phi(s)$, μ and ϕ for CUB and NLCUB models. In the CUB models, the transition probabilities are constant over t, s and given by

$$\phi_t(s) = 1 - \xi \quad \forall t, s. \tag{4}$$

with $s = 1, \ldots, m-1$, $t = 1, \ldots, m-1$ and $\phi_0 = \phi_0(1) := 1 - \xi$. In addition, we also have $\phi = \phi_t(s) = 1 - \xi$ and $\mu = (m-1)(1-\xi)$. In other words, in the CUB models $1 - \xi$, that is the *feeling* parameter, indicates the probability of increasing one rating point in one step of the feeling path.

In the NLCUB models, the transition probabilities result

$$\phi_t(s) = (1-\xi)\frac{\binom{t}{w_{g_s s}}(1-\xi)^{w_{g_s s}}\xi^{t-w_{g_s s}}}{\sum_{h=1}^{g_s}\binom{t}{w_{hs}}(1-\xi)^{w_{hs}}\xi^{t-w_{hs}}} \tag{5}$$

where $w_{hs} = y_{hs} - 1$ and with $s = 1, \ldots, m-1$, $w_{1s} \leq t < T$, $\phi_0 = \phi_0(1) := 0$ if $g_1 > 1$ and $\phi_0 = \phi_0(1) := 1 - \xi$ if $g_1 = 1$. When $T = m - 1$ and $g_s = 1$ for all s, formulas (4) and (5) coincide. In NLCUB, the expected number of one-rating-point increments during the feeling path is given by

$$\mu = \phi_0 + (1-\xi) \sum_{t=1}^{T-1} \sum_{s=1}^{m-1} \binom{t}{w_{g_s s}} (1-\xi)^{w_{g_s s}} \xi^{t - w_{g_s s}} \tag{6}$$

and $1 + \mu$ is the expected rating of the feeling approach, without the effect of the uncertainty approach.

2.3. Parameter estimation of the Nonlinear CUB models

In this paragraph we briefly recall the procedure proposed to estimate the parameters of a NLCUB model. It's worth pointing out that estimating a NLCUB model implies estimating both the parameters π, ξ and the parameters $g_1, \ldots, g_m$ describing the function l. Therefore, the transition probabilities and the shape of the transition plot are estimated from the data.

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 $\mathbf{g} = (g_1, \ldots, g_m)$ can be written as

$$L(\xi, \pi | \mathbf{g}; \mathbf{s}) = \sum_{i=1}^{n} \log \left\{ \pi \left[\sum_{h=1}^{g_{s_i}} \binom{T}{w_{hs_i}} (1-\xi)^{w_{hs_i}} \xi^{T - w_{hs_i}} \right] + (1-\pi) \frac{1}{m} \right\} \tag{7}$$

with $T = g_1 + \cdots + g_m - 1$. We obtain the estimates $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\pi}, \hat{\mathbf{g}})$ by the following procedure:

- fix a maximum value T_{max} for T;
- considering all the possible configurations of $g_1, \ldots, g_m$ such that $g_1 + \cdots + g_m \leq T_{max} + 1$, compute

$$\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_m) = \arg\max_{\mathbf{g}} \left\{ \max_{\xi, \pi} L(\xi, \pi | \mathbf{g}; \mathbf{s}) \right\};$$

- maximize (7) with respect to ξ and π to get

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

The number of possible configurations of $g_1, \ldots, g_m$ to be considered in the estimation procedure clearly depends on the values of m and T_{max}. For example, for $m = 5$ and $T_{max} = 8, 9, 10, 11$ we have $126, 252, 462, 687$ possible configurations of

$g_1, \ldots, g_m$, respectively. Estimation, along with other inferential issues, are the main challenges of the NLCUB models and further research is being devoted to refine some points, as discussed in Manisera and Zuccolotto (2014). With reference to the choice of T_{max}, which could appear discretionary, some considerations are discussed in Subsection 4.3 of this paper.

3. Linear and nonlinear decision processes

The decision process underlying the individuals' responses on a rating scale has been defined to be linear or nonlinear according to whether the transition probabilities $\phi_t(s)$ are constant on non-constant for different t and s. This implies that, for linear processes, the transition plot shows a straight line, since the probability of increasing one rating point in the next step of the feeling path is constant for every rating in every step (as in the example of Figure 1, left).

Manisera and Zuccolotto (2014) derive, under some general assumptions, a sufficient condition for linearity and show that CUB (*i*) is a particular case of the general framework and (*ii*) meets the sufficient condition for linearity. The NLCUB models, instead, are a nonlinear variant of the general model and this is a reason for their name. A graphical explanation is also possible, since the transition plot of the NLCUB models generally shows a nonlinear broken line, giving interesting insights on the way the respondents perceive the response scale and, in particular, the distance among the response categories (as in the example of Figure 1, right).

Starting from the above definition of linearity, in this paper we propose to measure the degree of nonlinearity expressed by a NLCUB model as the standard deviation of the transition probabilities. Formally, we define the following nonlinearity index:

$$\lambda(\xi, \mathbf{g}) = \sigma(\phi_t(s)) / \max(\sigma) \tag{8}$$

where $\sigma(\phi_t(s))$ is the standard deviation of $\phi_t(s)$, $\forall t, s \in \Phi$ with Φ denoting the set containing all the pairs $(t, s) : \exists\ \phi_t(s)$. The value $\max(\sigma)$ can be obtained as follows. Let $|\Phi| = k$, where $|\cdot|$ denotes the cardinality of a set. For odd k, $\max(\sigma)$ is the value of $\sigma(\phi_t(s))$ in the extreme situation where $k/2$ probabilities $\phi_t(s)$ equal 0 and the remaining $k/2$ probabilities equal 1; in this case with simple algebra we obtain $\sigma(\phi_t(s)) = \sqrt{1/2 - 1/4} = \sqrt{1/4}$. For even k, $\max(\sigma)$ is reached when either $(k-1)/2$ probabilities $\phi_t(s)$ equal 0 and the remaining $(k+1)/2$ probabilities equal 1 or $(k+1)/2$ probabilities $\phi_t(s)$ equal 0 and the remaining $(k-1)/2$ probabilities equal 1. In these two cases we have $\sigma(\phi_t(s)) = \sqrt{(k+1)/2k - (k+1)^2/4k^2}$ and $\sigma(\phi_t(s)) = \sqrt{(k-1)/2k - (k-1)^2/4k^2}$, respectively. In both cases we finally obtain $\sigma(\phi_t(s)) = \sqrt{1/4 - 1/4k^2)}$. Therefore, we have

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

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 respect to its maximum. The nonlinearity index λ is expressed as a function of $(\xi, \mathbf{g})$ and does not depend on π, because the transition probabilities only pertain the feeling approach.

4. Stylized facts

In this section, we empirically observe the statistical features of several different NLCUB models, obtained by varying the parameters in the parameter space so as to systematically explore a huge number of possible combinations. In the end, we draw some stylized facts considering the following issues:

- the extrapolation of some particular cases concerning the existence of linear DPs within the NLCUB framework;
- the analysis of some evidence about the nonlinearity pattern of different NLCUB models;
- the possibility to draw some remarks about the choice of T_{max} for estimation purposes.

In order to gain insights about these issues, we have computed the values of $\phi_t(s)$, $\phi(s)$ $(t = 1, 2, \ldots, T-1; s = 1, \ldots, m-1)$ and μ for all the NLCUB models obtained crossing the experimental conditions reported in Table 2 (altogether, 4,752 different combinations of m, T, ξ). For each case, we have investigated all the possible configurations of $(g_1, \ldots, g_m)$, so that a total number of 13,695,858 NLCUB models have been considered in this systematic study.

Table 2. Experimental conditions

Parameter	Explored values
m	$4, 5, 6, 7$
T	$m-1, m, m+1, \ldots, 3m-1$
ξ	$0.01, 0.02, \ldots, 0.98, 0.99$

For illustrative purposes, we report the transition plots for 36 selected combinations of m, T, ξ (Figures 2 - 5; for each case, all the combinations of $(g_1, \ldots, g_m)$). At a first sight we notice that (1) each combination seems to contain (at least) one linear transition plot, (2) the shapes of the transition plots tend to become "more nonlinear" with increasing values of T and decreasing values of ξ. These two rough remarks will be more deeply analysed in the next three subsections.

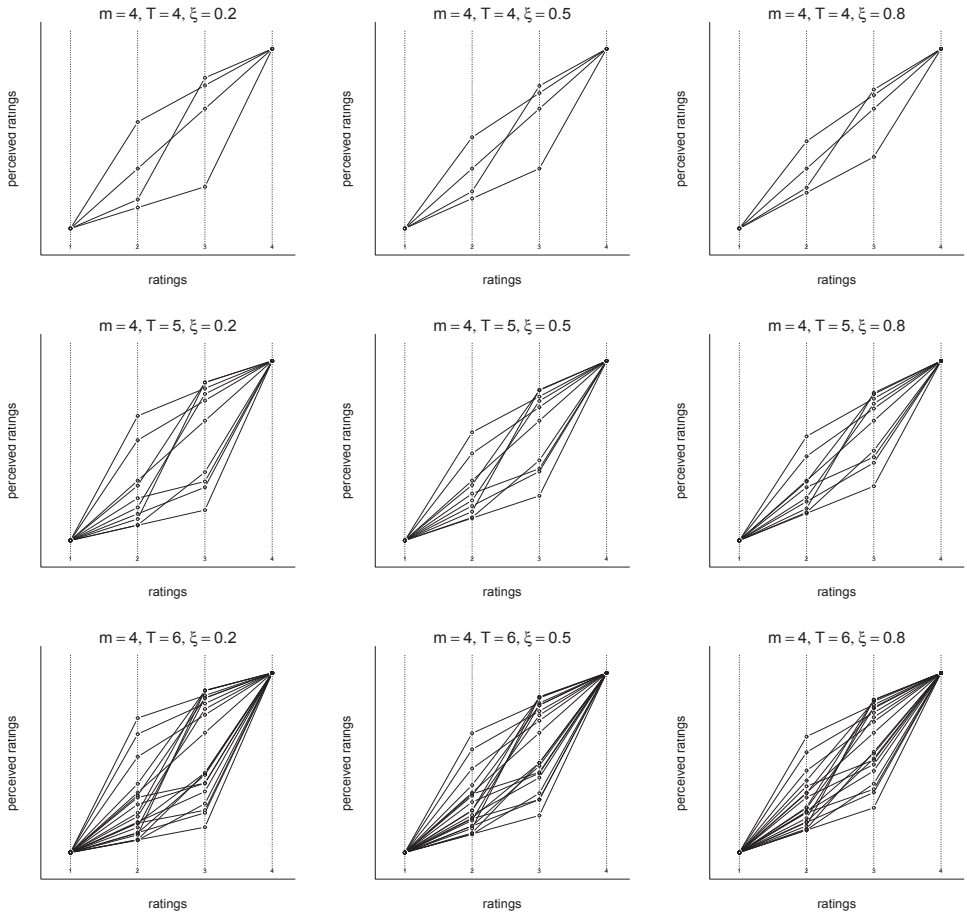

Figure 2. Transition plots for the cases with $m = 4$, $T = 4, 5, 6$ *(top, middle, bottom),* $\xi = 0.2, 0.5, 0.8$ *(left, middle, right)*

4.1. Linear DPs within the NLCUB framework

Within the NLCUB framework, the sufficient condition for linearity in Manisera and Zuccolotto (2014) is satisfied only by the configuration $g_1, \ldots, g_m = (1, \ldots, 1)$, that is, when the NLCUB collapses into a classical CUB model. However, we are aware that other linear DPs may exist within the NLCUB framework, as the above mentioned condition is not necessary. Our empirical investigation has found that, for each combination of m and T, a linear DP is generated by the configuration of $g_1, \ldots, g_m$ such that $g_s = 1$ with $s = 1, \ldots, m-1$ and $g_m = T - m + 2$, whatever the value of ξ. As for future

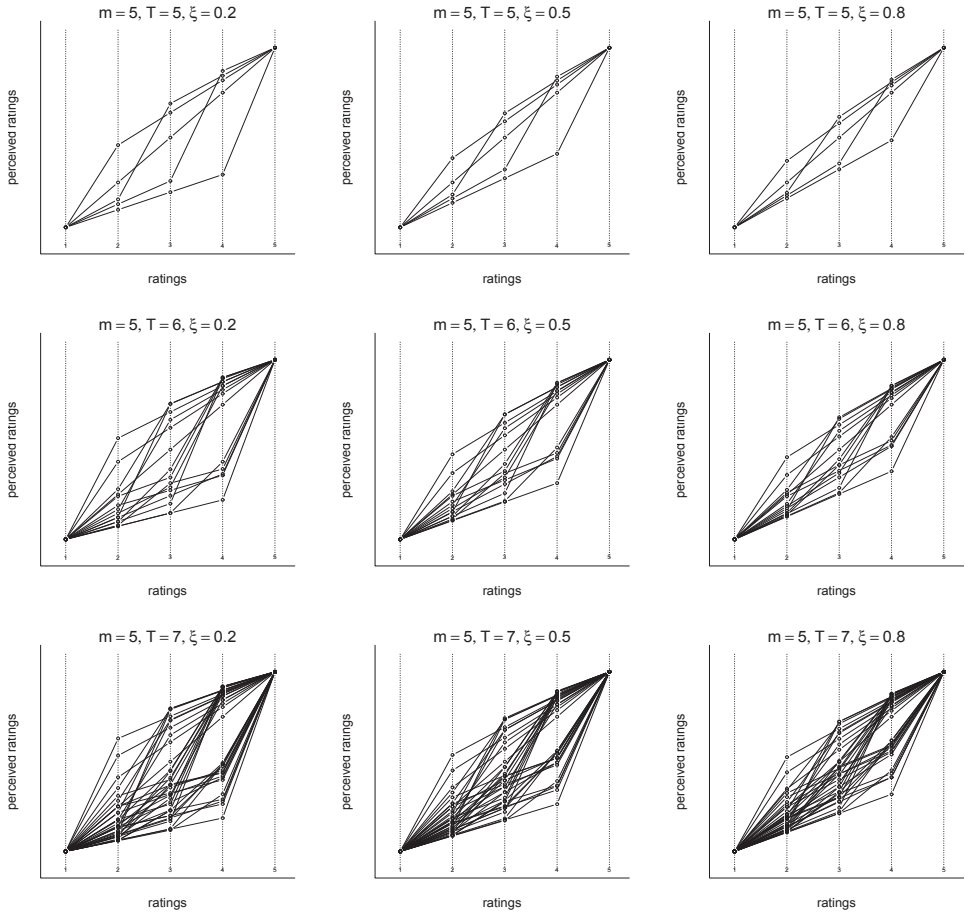

Figure 3. Transition plots for the cases with $m = 5$*,* $T = 5, 6, 7$ *(top, middle, bottom),* $\xi = 0.2, 0.5, 0.8$ *(left, middle, right)*

research, this constitutes a clear suggestion for trying to define a sufficient and necessary condition. Although all these different DPs meet the definition of linearity, their feeling paths work according to different mechanisms, so that the same values of ξ correspond to different values of μ (see Figure 6). In detail, with high values of T, μ tends to remain fixed at its highest value until ξ reaches a given threshold.

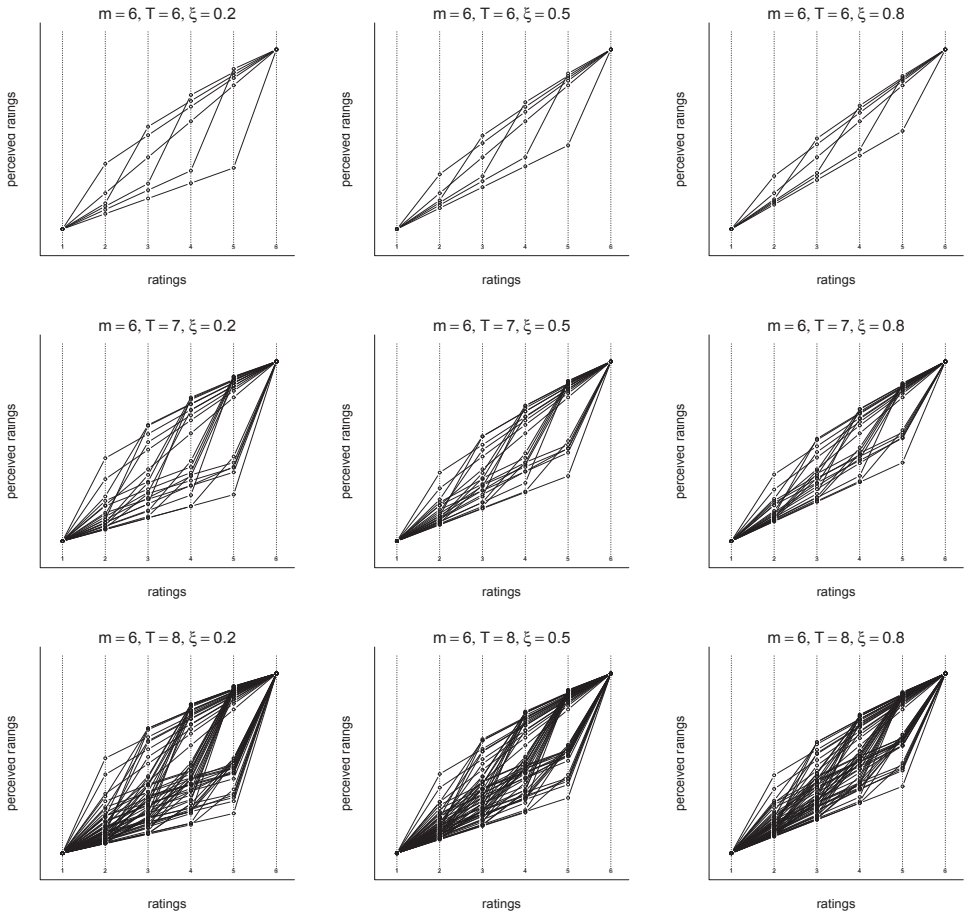

Figure 4. Transition plots for the cases with $m = 6$, $T = 6, 7, 8$ *(top, middle, bottom),* $\xi = 0.2, 0.5, 0.8$ *(left, middle, right)*

4.2. Empirical evidence about nonlinearity

In this subsection we explore the relationships between the nonlinearity index $\lambda(\xi, \mathbf{g})$ and the parameters ξ and T. Figure 7 shows the overall plot of $\lambda(\xi, \mathbf{g})$ versus ξ and the corresponding partial plots for some selected values of T, with $m = 4$, the remaining three cases ($m = 5, 6, 7$) being substantially similar.

The graphs clearly show that the highest levels of nonlinearity can be reached with low values of ξ, provided that T is large enough (from $T = 7$ onwards, in this case) and that higher values of T allow $\lambda(\xi, \mathbf{g})$ to cover a wider range of values. The linear

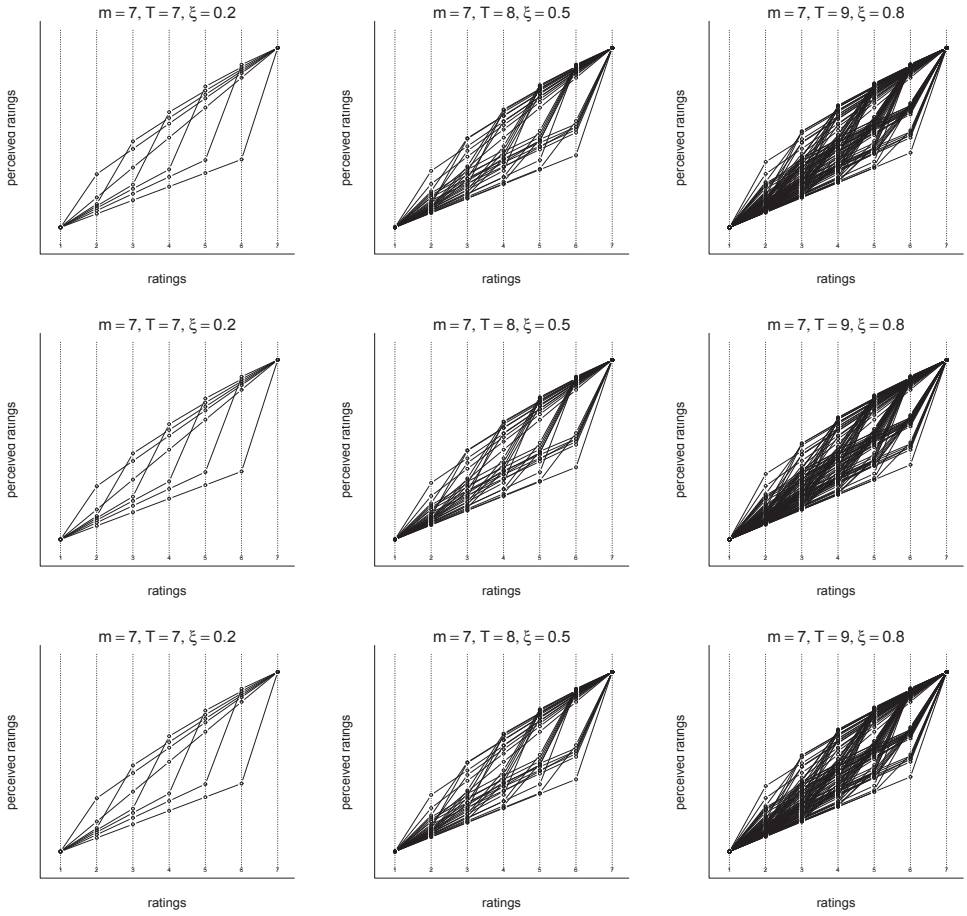

Figure 5. Transition plots for the cases with $m = 7$, $T = 7, 8, 9$ *(top, middle, bottom),* $\xi = 0.2, 0.5, 0.8$ *(left, middle, right)*

correlation between $\lambda(\xi, \mathbf{g})$ and ξ results -0.9206, -0.9529, -0.9664, -0.9754 for $m = 4, 5, 6, 7$, respectively. The relationship between $\lambda(\xi, \mathbf{g})$ and T can be evaluated by inspecting the boxplots in Figure 8, showing that when m increases, we need higher and higher values of T to reach the maximum level of nonlinearity. On the other hand, the median values of $\lambda(\xi, \mathbf{g})$ seem to increase very slowly after the first values of T.

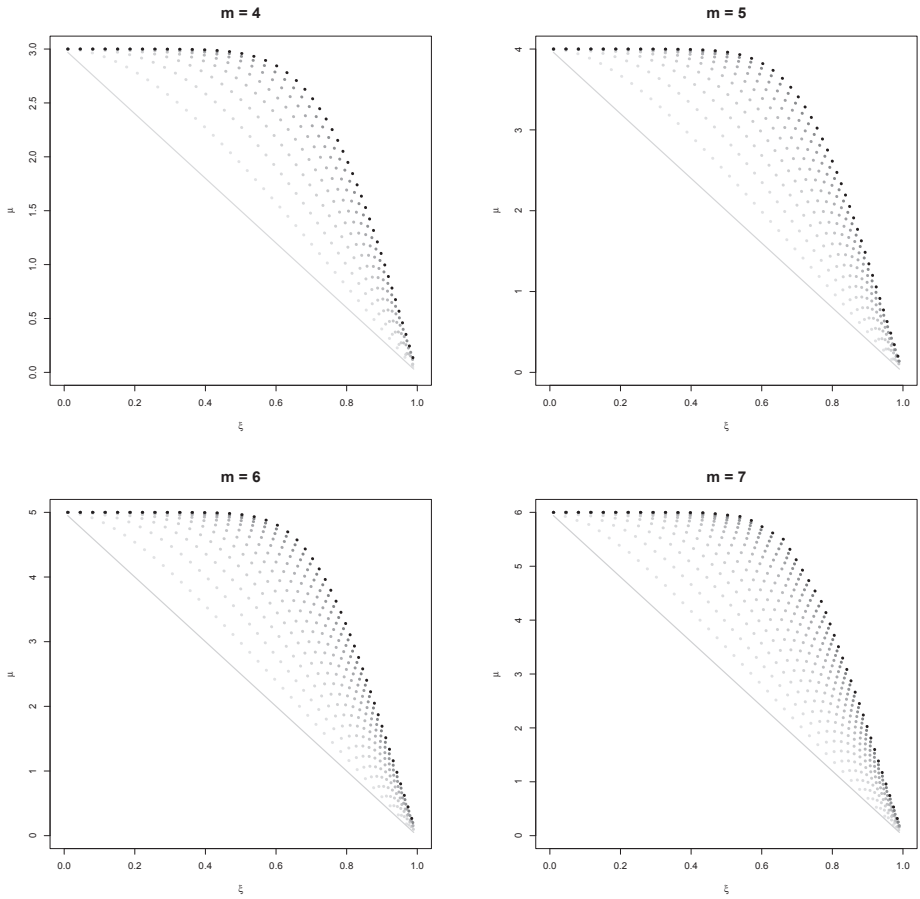

Figure 6. Relationship between ξ and μ (gray-level scale proportional to the values of T; solid line denoting the configuration with $T = m - 1$, i.e. the CUB model) for $m = 4, 5, 6, 7$

4.3. Remarks about the choice of T_{max}

The estimation procedure for the NLCUB models requires the definition of the maximum value T_{max} of T. This choice is rather crucial as high values of T may cause both identifiability and overfitting problems (Manisera and Zuccolotto, 2014), which can be kept under control by forcing T within a given range. The choice of a relatively small value for T_{max} is also justified from the point of view of the unconscious DP, since the commitment of respondents in formulating judgments is generally moderate, so it is rea-

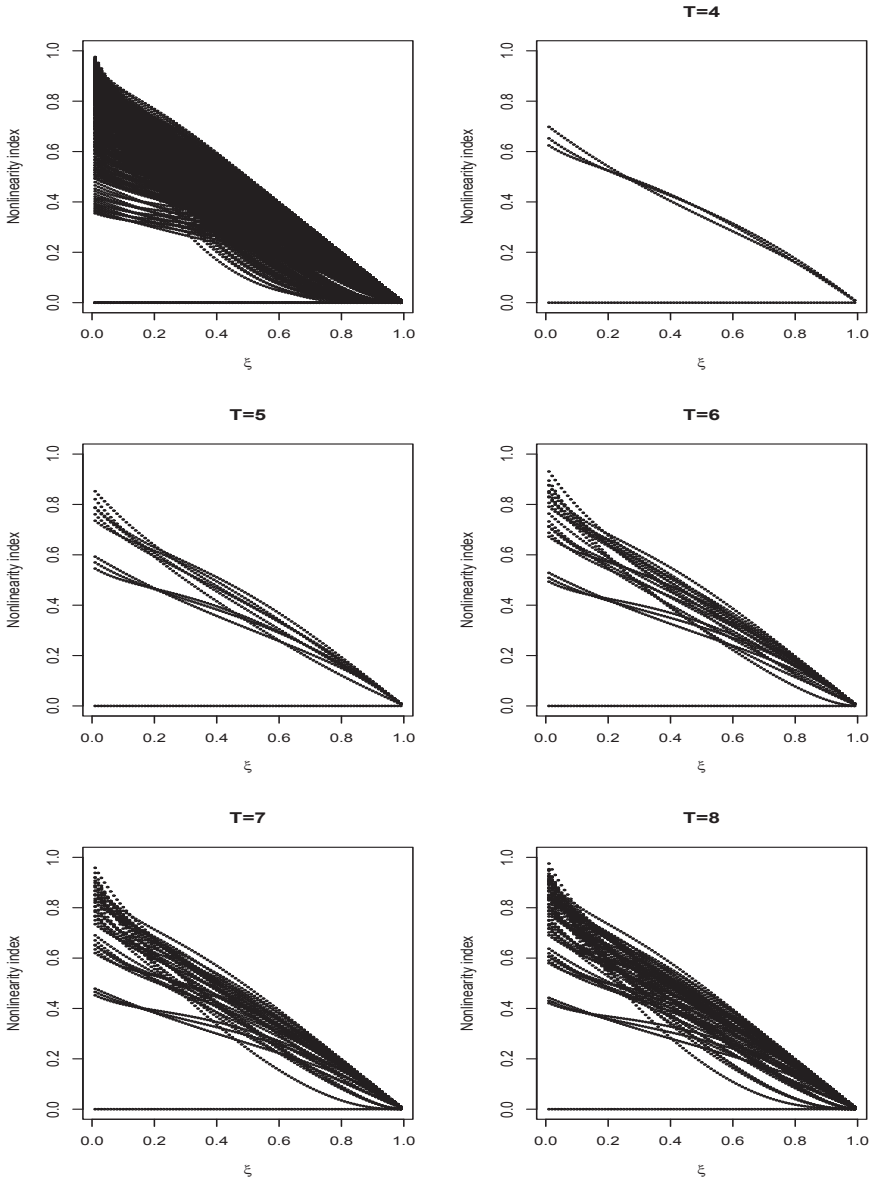

Figure 7. Overall plot of $\lambda(\xi, \mathbf{g})$ versus ξ (top left panel) and partial plots for some selected values of T ($m = 4$)

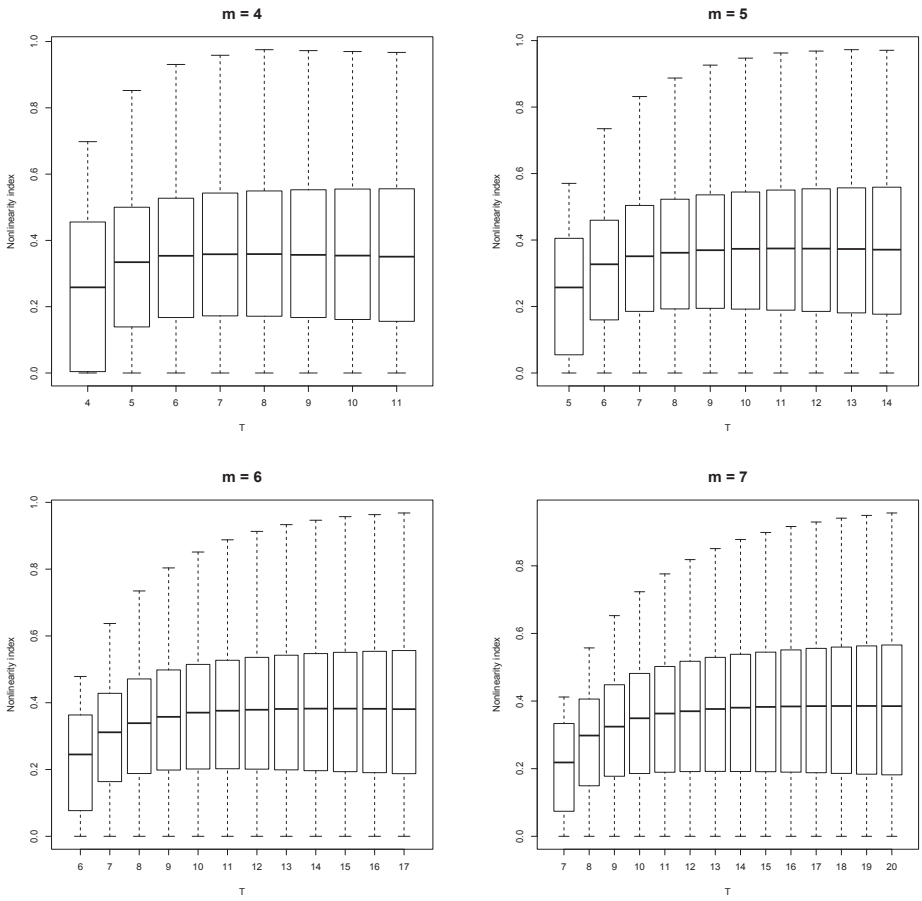

Figure 8. Boxplots of $\lambda(\xi, \mathbf{g})$ *given* T *for* $m = 4, 5, 6, 7$

sonable to assume a limited number of steps in the feeling path, whatever the complexity of the evaluated item.

The results of the systematic analysis carried out in this work can provide some useful suggestions about the choice of T_{max}. In summary, we have to balance two opposite needs:

- to define a model with flexibility enough to reproduce several nonlinear patterns: this requires to fix a high value for T_{max}, which allows the nonlinearity index $\lambda(\xi, \mathbf{g})$ to cover a wider range of values, as pointed out in Subsection 4.2;
- to pay attention on identifiability and overfitting problems: this requires to fix a

low value for T_{max}.

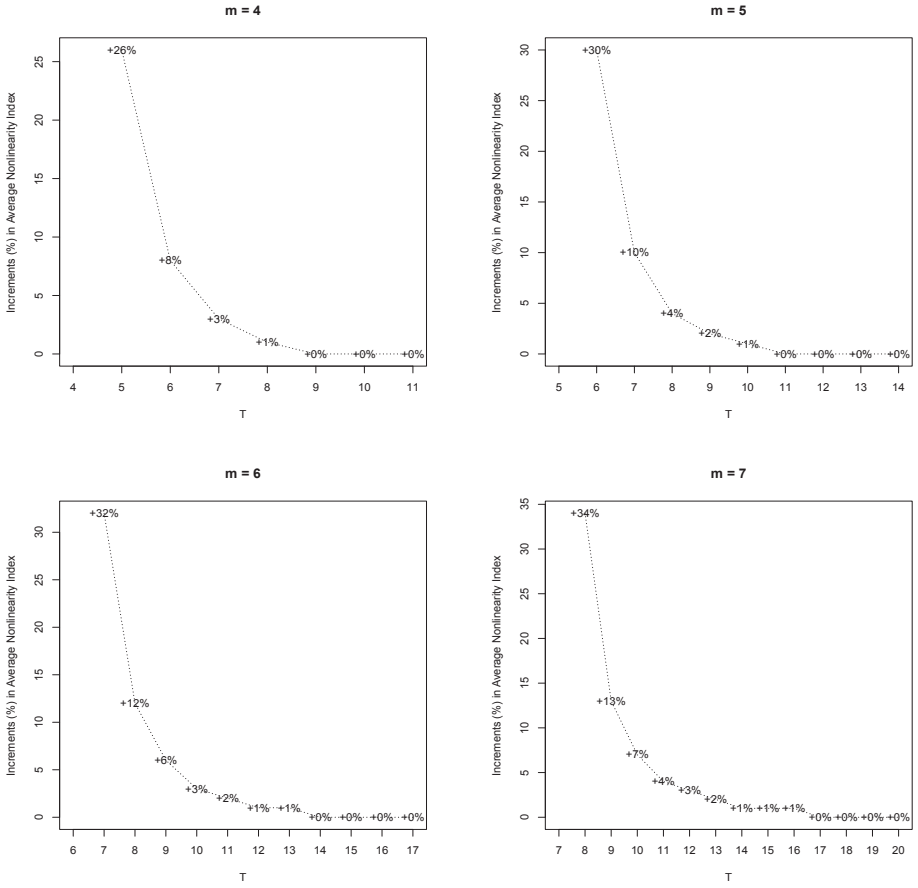

Figure 9. Increments in the average of $\lambda(\xi, \mathbf{g})$ given T, when moving from $T-1$ to T, for $m = 4, 5, 6, 7$

Figure 9 shows how the average of $\lambda(\xi, \mathbf{g})$ given T increases when moving from $T-1$ to T. We immediately note that the increments tend to be negligible after the first values of T. We feel that a good balance between the above mentioned opposite needs can be found when T_{max} approximately equals $2m$. In addition, if we establish that NLCUB models should be flexible enough to guarantee a nonlinearity index with an acceptably large range (from 0 to, at least, 85-95%) for varying m, $T_{max} = 2m - 1$ seems preferable (Table 3). Being aware that the choice of T_{max} is, to a certain extent,

discretionary, we are convinced that T_{max} could be conveniently be set at $2m - 1$. This corresponds to select a Shifted Binomial random variable with twice the categories of the response scale. Additionally, although the estimation procedure of the NLCUB model is fairly not time-consuming, this choice allows to keep the number of possible configurations of $g_1, \ldots, g_m$ reasonably low, so reducing computational time.

Table 3. Values of the nonlinearity index (in %) for some m and T; only the values in [85%,95%] are displayed

T	$m = 4$	$m = 5$	$m = 6$	$m = 7$
$2m - 3$	85			
$2m - 2$	93	89	85	
$2m - 1$	95	93	89	85
$2m$		95	91	88
$2m + 1$			93	90
$2m + 2$			94	92
$2m + 3$			95	93
$2m + 4$				94
$2m + 5$				95
$2m + 6$				95

5. Conclusions

In this paper we have presented a systematic analysis of some main features of the Nonlinear CUB models (NLCUB), aimed at giving insights on the behaviour of this new class of models and suggestions about the future theoretical developments.

In detail, we have explored three issues, concerned with (1) the existence of linear DPs within the NLCUB framework, (2) the nonlinearity patterns expressed by different NLCUB models, (3) the choice of the value T_{max} in the estimation procedure.

The computational method exploited in this study was not, as usual, simulation. In fact, we have derived the theoretical statistical features of all the NLCUB models obtained by varying the parameters values, so that over 13 millions different models have been included in the analysis.

About point (1), we have demonstrated that the CUB case is not the unique linear DP within the NLCUB framework, thus confirming the need of devoting future research to the definition of a sufficient and necessary condition for linearity, whose possible formulation can be conjectured relying on the presented empirical evidences.

Points (2) and (3) are strictly connected each other. In fact, we have found that both the parameters T and ξ play an important role in the nonlinearity pattern of the NLCUB models. This evidence, although being of limited importance with reference to ξ, whose

parameter space is restricted to $[0, 1]$, is very meaningful for what concerns T. It is then able to give some suggestions about the choice of T_{max} in the estimation procedure (at least for the explored values of m, which are, however, the most common ones in real situations).

Starting from this systematic study, future research can be devoted to derive a theo-retical formalization of the obtained empirical evidence.

Acknowledgements
Research funded by STAR project (University of Naples Federico II - CUP: E68C13000020003).

References

Agresti, A. (2013). *Categorical Data Analysis*, 3^{rd} edition, J. Wiley & Sons, New York.

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

Cerchiello, P., Iannario, M., Piccolo, D. (2010). Assessing risk perception by means of ordinal models, in: M. Corazza, C. Pizzi (eds.): *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, Springer-Verlag, pp.75–83.

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

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

Iannario M. (2007) Dummy variables in CUB models, *Statistica*, LXVIII, 2.

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

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

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

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

Iannario, M. (2012b). Hierarchical CUB Models for Ordinal Variables, *Communication in Statistics - Theory and Methods*, **41**, 3110–3125.

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. (2014). A Short Guide to CUB 3.0 Program. Available at https://www.researchgate.net/publication/260959050.

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

Iannario, M., Piccolo, D. (2012). CUB Models: Statistical Methods and Empirical Evidence, in: R. S. Kenett, S. Salini (eds.): *Modern Analysis of Customer Surveys*, NY: Wiley, pp. 231–258.

Manisera, M., Zuccolotto, P. (2013). Nonlinear CUB models. In: T. Minerva, I. Morlini, F. Palumbo (eds.): *Book of Abstracts Cladag 2013, 9th Meeting of the Classification and Data Analysis Group*, CLEUP, pp. 288–291.

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

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

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

Piccolo, D. (2013). Inferential issues on CUBE models with covariates, *Communication in Statistics - Theory and Methods*, in press.

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

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