A Mixture Model for the Analysis of Categorical Variables Measured on Five-point Semantic Differential Scales

Marica Manisera ⬤
Big&Open Data Innovation
Laboratory (BODaI-Lab)
University of Brescia
Italy

Manlio Migliorati ⬤
Big&Open Data Innovation
Laboratory (BODaI-Lab)
University of Brescia
Italy

Matteo Ventura ⬤
Big&Open Data Innovation
Laboratory (BODaI-Lab)
University of Brescia
Italy

Paola Zuccolotto ⬤
Big&Open Data Innovation
Laboratory (BODaI-Lab)
University of Brescia
Italy

Abstract

Ordered response scales are often used in questionnaires to measure individuals' attitudes or perceptions. Among different response scale formats, we focus on multi-point semantic differential scales, requiring the respondent to position himself/herself on a rating between two bipolar adjectives. The obtained rating data require appropriate statistical models. We resort to the CUM model (Combination of a discrete Uniform and a - linearly transformed - Multinomial random variable), recently proposed in the framework of the CUB (Combination of discrete Uniform and shifted Binomial random variables) class of models. CUM is also suited to all the ordinal response scales with a middle "indifference" option. In the seminal paper on CUM, the methodological approach was developed for an odd number m of response categories, while simulations, case studies and implementation in R were limited to $m = 7$. The objective of this paper is to extend the original proposal and investigate the model performance in the case of $m = 5$, which often arises in real situations. The R functions for fitting a CUM model with $m = 5$ are implemented and made available; simulation studies are developed and compared with results obtained for $m = 7$ and a case study concerned with the evaluation of museums' visitor experience is proposed.

Keywords: categorical ordinal variables, rating data, multinomial distribution, CUM model, CUB models.

1. Introduction

Ordered response scales are often used in questionnaires in order to measure individuals' perceptions, attitudes, evaluations, habits, behaviours. For example, when we investigate patient satisfaction, respondents are often asked to rate their satisfaction, with a given service, on a response scale from 1="completely dissatisfied" to, say, 5="completely satisfied." Among a variety of possible response scales (Dawis 1987), we focus on (multi-point) semantic differential scales (mSDS), used when respondents are asked to assess their position between two opposite words/concepts/adjectives (for example, sad and happy, weak and strong, dissatisfied and satisfied, etc.). Instead of positioning themselves on a continuous trait, for example by drawing a tick on a segment, when a semantic differential scale is multi-rating, respondents select their favourite option among multiple ratings, from one to the other of the two words/concepts/adjectives.

When dealing this kind of response scales, the middle option plays a relevant role, because it indicates a central position between two extremes. However, this can be considered true in all the response scales having an odd number of ordered categories with a midpoint that means indifference, like "neutral," "neither positive nor negative," or "neither agree nor disagree", as often is the case of the Likert scales commonly used in several fields.

Rating data obtained by the administration of questionnaires need to be modelled with appropriate statistical models and methods, properly designed to account for the categorical ordinal nature of these variables (Agresti 2010, 2013). In addition, specific models can be considered when rating data come from mSDS.

Among the statistical models existing in the literature for the analysis of ordinal data measured on semantic differential scales (SDS), we focus on a mixture model, called CUM model (Combination of a discrete Uniform and a - linearly transformed - Multinomial random variable), recently proposed by Manisera and Zuccolotto (2022) in the framework of the CUB class of models (where CUB stands for Combination of discrete Uniform and shifted Binomial random variables) .

The CUB class of models (Piccolo 2003; D'Elia and Piccolo 2005) innovates with respect to other existing models for categorical data by interpreting the decision process that underlies the respondent's choice as the result of a combination of two latent components. They are called *feeling* and *uncertainty* and measure (i) the individuals' level of perception, attitude, evaluation (for example, satisfaction) with the object under evaluation and (ii) the indecision that surrounds any choice among multiple ratings. The two components are then collected in a mixture distribution, combining a (Shifted) Binomial random variable (r.v.) that models feeling and a Uniform r.v. that models uncertainty. CUB models have been extended in several directions and applied in many fields (Piccolo and Simone 2019) . The most recent contributions deal with the possibility to model subjective heterogeneity Simone, Tutz, and Iannario (2020), some computational issues (Simone 2021; Cerulli, Simone, Di Iorio, Piccolo, and Baum 2022), a new approach combining CUB models to decision trees (Cappelli, Simone, and Iorio 2019; Simone, Cappelli, and Di Iorio 2019) and cluster analysis (Biasetton, Disegna, Barzizza, and Salmaso 2023).

The CUM model was introduced (Manisera and Zuccolotto 2022) to extend the CUB paradigm to specifically model the peculiar decision process followed by respondents facing mSDS (or rating scales with a neutral midpoint). CUM was developed in the framework modelling the decision process in Manisera and Zuccolotto (2014a) as a mixture model combining a linearly transformed Multinomial (ltM) and a discrete Uniform r.v. In Manisera and Zuccolotto (2022) the methodological approach of the CUM model was developed for the general situation of an odd number m of response categories, while simulations, case studies and implementation in R (R Core Team 2019) were limited to $m = 7$.

The aim of this paper is to further investigate the functioning of the CUM model in the presence of a semantic differential response scale with $m = 5$ categories. In particular, we develop a simulation study following the configuration in Manisera and Zuccolotto (2022)

to compare results of $m = 5$ with those of $m = 7$, propose a case study and adapt the R functions to cope with $m = 5$.

At `https://bodai.unibs.it/cub/`, interested readers can find references, the R functions and some data examples about the CUM model, besides other interesting methodological advancements developed by the "Big & Open Data Innovation Laboratory" research group of the University of Brescia, Italy on rating data analysis.

The paper is structured as follows. The CUM model is briefly recalled in Section 2. Subsection 2.1 declines some of the CUM characteristics in the case of $m = 5$. Section 3 presents the results of the simulation study, focusing the attention on the comparison between the two situations of $m = 7$ and $m = 5$. In Section 4 we present an application of the CUM model to real data coming from a recent survey aimed at evaluating the museums' visitor experience. Section 5 draws some conclusions.

2. The CUM model

According to Manisera and Zuccolotto (2014a), the decision process (DP) underlying the respondents' choice on a response scale with ordered categories is composed of two components that, borrowing the CUB terminology, are called feeling and uncertainty. They originate in the respondent's mind, where two mind approaches are unconsciously combined to finally obtain a final choice on the response scale.

Feeling and uncertainty approaches acting in the DP are rigorously defined in Manisera and Zuccolotto (2014a). Several variants of the CUB models can be derived from the DP as special cases, for example the NonLinear CUB model (Manisera and Zuccolotto 2014a, 2015, 2014b). In addition, novel models can be proposed by modifying some of the ingredients of the DP (distribution of the r.v.s involved, and the so-called accumulation and likertization functions) (Manisera and Zuccolotto 2019), like the CUM model (Manisera and Zuccolotto 2022).

In the CUM model, the observed rating r, $r = 1, 2, \ldots, m$, with $m = 2k+1$, is the realization of R, a discrete r.v. with probability mass given by:

$$P(R = r|\pi, \xi_D, \xi_U) = \pi W_k(r|\xi_D, \xi_U) + (1 - \pi)U_m, \tag{1}$$

where $\xi_D \in [0, 1]$, $\xi_U \in [0, 1]$, $\pi \in (0, 1)$. For a given m, R is distributed as a mixture of a ltM r.v. $W_k(r|\xi_D, \xi_U)$ and a discrete Uniform r.v. U_m defined over $\{1, 2, \ldots, m\}$.

In detail, U_m models the uncertainty component and $1 - \pi$ weighs the uncertainty. Therefore, $1 - \pi$ can be then assumed as a measure of the uncertainty component in the mixture.

On the feeling side, the r.v. $W_k(r|\xi_D, \xi_U) = P(W_k = r|\xi_D, \xi_U)$ is defined as

$$W_k = [-1 \quad 1 \quad 0]\mathbf{M}_k + k + 1, \tag{2}$$

that is a linear transformation of the Multinomial r.v. $\mathbf{M}_k = \mathbf{M}_k(k_1, k_2)$ with trial parameter $k = (m - 1)/2$ and taking values $(k_1, k_2, k - k_1 - k_2)$.

The Multinomial r.v. $\mathbf{M}_k$ originates from a DP where the feeling path unconsciously followed by the respondent is composed of k steps. At each step, the respondent formulates an elementary judgement that can be modelled by a Multinoulli distribution assuming value $[0, 1, 0]$ with probability ξ_U when the respondent makes one step towards the highest rating (the right-hand side of the response scale), value $[1, 0, 0]$ with probability ξ_D when the respondent makes one step towards the lowest rating (the left-hand side of the response scale), value $[0, 0, 1]$ with probability $1 - \xi_D - \xi_U$ when the respondent considers to stay still. The judgement achieved at the k-th step of the feeling path is given by the *total* number of elementary judgements provided up to the k-th step, where $M_{k,D}$ and $M_{k,U}$ are respectively the total number of steps towards the lowest and highest rating made up to the k-th step. Therefore, the total number of elementary judgements at step k can be modelled by the sum

of independent Multinoulli r.v.s, that is a Multinomial r.v. $\mathbf{M}_k = [M_{k,D}, M_{k,U}, M_k]$ with trial parameter k. The linear transformation (2) is then needed to map from the domain of $\mathbf{M}_k$ to $\{1, 2, \ldots, m\}$.

The CUM model has then two feeling parameters, ξ_D and ξ_U, which are a measure of the respondents' propensity to move towards the left-hand side (i.e., the lowest rating) and the right-hand side (i.e., the highest rating) of the response scale, respectively, starting from the the mid-point $k + 1$. Then, the difference $1 - \xi_D - \xi_U$ measures the propensity of staying still.

In Manisera and Zuccolotto (2022), the probability mass function of $W_k(r|\xi_D, \xi_U)$ was derived in the general case of $m = 2k + 1$ as follows:

$$p(W_k = r = k + h) = \begin{cases} \mathbf{M}_k\left(\max(-2\alpha, 0), \max(2\alpha, 0)\right) + \\ +\mathbf{M}_k\left(\max(-2\alpha + 1, 1), \max(2\alpha + 1, 1)\right) + \cdots + \\ +\mathbf{M}_k\left(d(k/2) - \alpha, d(k/2) + \alpha\right) \\ \text{if } h = 2\alpha + 1 \\ \\ \mathbf{M}_k\left(\max(-2\alpha + 1, 0), \max(2\alpha - 1, 0)\right) + \\ +\mathbf{M}_k\left(\max(-2\alpha + 2, 1), \max(2\alpha, 1)\right) + \cdots + \\ +\mathbf{M}_k\left(u(k/2) - \alpha, u(k/2) + \alpha - 1\right) \\ \text{if } h = 2\alpha \end{cases} \tag{3}$$

where $h \in \{-k + 1, -k + 2, \cdots, k + 1\}$, $\alpha \in \{-d(k/2), -d(k/2) + 1, \cdots, u(k/2)\}$, $d(\cdot)$ and $u(\cdot)$ round down and up, and $\mathbf{M}_k(k_1, k_2)$ is the probability of a Multinomial r.v. (with k trials) taking values equal to $(k_1, k_2, k - k_1 - k_2)$.

Formula (3) is the starting point to derive, in Subsection 2.1, the probability mass function of $W_2(r|\xi_D, \xi_U)$, that is $W_k(r|\xi_D, \xi_U)$ for $m = 5$ (and so $k = 2$).

Parameter estimation was carried out by Maximum Likelihood method. For fixed m and given the ratings $\mathbf{r} = (r_1, \ldots, r_n)'$ of n independent subjects in a sample, the log-likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{r}) = \sum_{i=1}^{n} \log[p(R = r_i|\boldsymbol{\theta})] = \sum_{r=1}^{m} n_r \log p_r, \tag{4}$$

where $\boldsymbol{\theta} = (\pi, \xi_D, \xi_U)'$, n_r is the (absolute) frequency of r, and $p_r = p(R = r|\boldsymbol{\theta})$.

Since the maximization of the log-likelihood function by numerical methods (Nelder and Mead 1965) showed some convergence issues and high sensitivity to starting values, parameters of the CUM models can be estimated by resorting to the EM algorithm (Manisera and Zuccolotto 2022), as is commonly done with mixture models and, in particular, in mixture distributions of the CUB class (Manisera and Zuccolotto 2017).

The derivatives to be used to compute the M-step of the EM algorithm for CUM estimation and the (asymptotic) variance-covariance matrix were obtained in (Manisera and Zuccolotto 2022) and there derived for $m = 7$. In this paper, we will use the same approach and with simple algebra we will adapt the computation of those derivatives to the case of $m = 5$ (computations available upon request to the Authors).

The CUM model has a complex likelihood surface which causes stability issues in the estimates via EM algorithm. In addition, the identifiability of CUM model over the whole parameter space is still an open issue. Therefore, for estimation in this paper we checked for uniqueness of the EM estimates obtained from a grid of different starting values.

Parameter estimates can be effectively represented with nice ternary plots (Manisera and Zuccolotto 2022). The dissimilarity (*diss*) index, as reported in (Piccolo and Simone 2019), page 408, is a measure of the distance between the observed relative frequencies and the estimated probabilities; it is a normalized index ($diss \in [0, 1]$) and can be used to assess the goodness of fit of a CUM model, with lower values indicating a better fit.

2.1. Characteristics of the CUM model for five-point semantic differential scale

We now consider a five-point SDS, where 1 and 5 indicate two opposite adjectives, for example sad and happy or completely dissatisfied and completely satisfied. First, if the rating scale has 5 categories, the feeling path has 2 steps, and starting from the middle option 3, the total number of steps $M_{k,D}$ towards the lowest rating can be 0, 1 or 2; the total number of steps $M_{k,U}$ towards the highest rating can be 0, 1 or 2. The linear transformation in (2) allows to obtain the ratings $1, ..., 5$ as in Table 1, that reports the values of the linearly transformed r.v. W_2 for the realizations of $M_{2,D}$ and $M_{2,U}$.

Table 1: Values assumed by the r.v. W_2 for the possible realizations of the Multinoulli r.v.s $M_{2,D}$ and $M_{2,U}$

W_2		$M_{2,U}$		
		0	1	2
$M_{2,D}$	0	3	4	5
	1	2	3	-
	2	1	-	-

Second, following Formula (3) applied to case $m = 5$, we have $h \in [-2+1, ..., 2+1] = [-1, 0, 1, 2, 3]$, $\alpha \in [-1, 0, 1]$, and we can obtain the probability mass function of W_2, which is given by:

$$P(W_2 = k + h = 2 + -1 = 1) = \mathbf{M_2}(2, 0) \qquad h = -1, \alpha = -1$$

$$P(W_2 = k + h = 2 + 0 = 2) = \mathbf{M_2}(1, 0) \qquad h = 0, \alpha = 0$$

$$P(W_2 = k + h = 2 + 1 = 3) = \mathbf{M_2}(0, 0) + \mathbf{M_2}(1, 1) \qquad h = 1, \alpha = 0$$

$$P(W_2 = k + h = 2 + 2 = 4) = \mathbf{M_2}(0, 1) \qquad h = 2, \alpha = 1$$

$$P(W_2 = k + h = 2 + 3 = 5) = \mathbf{M_2}(0, 2) \qquad h = 3, \alpha = 1$$

3. Simulation study

In this section, we show the results from a simulation study performed fitting the CUM model with $m = 5$ (hereafter CUM5) to data generated on the basis of the parameter values summarised in Table 2. The 18 scenarios are the same of the simulation study performed to evaluate the CUM model with $m = 7$ (CUM7) in Manisera and Zuccolotto (2022). For each scenario $iter = 1000$ simulations with $n = 1000$ observations are executed.

Table 2: Summary of parameter values used in the simulation study

	a			**b**			**c**		
	π	ξ_D	ξ_U	π	ξ_D	ξ_U	π	ξ_D	ξ_U
Case 1	0.3	0.2	0.1	0.5	0.2	0.1	0.7	0.2	0.1
Case 2	0.3	0.5	0.2	0.5	0.5	0.2	0.7	0.5	0.2
Case 3	0.3	0.8	0.1	0.5	0.8	0.1	0.7	0.8	0.1
Case 4	0.3	0.2	0.4	0.5	0.2	0.4	0.7	0.2	0.4
Case 5	0.3	0.4	0.5	0.5	0.4	0.5	0.7	0.4	0.5
Case 6	0.3	0.1	0.7	0.5	0.1	0.7	0.7	0.1	0.7

3.1. CUM5 simulation study

The ternary plots of the CUM5 simulation study for the 18 scenarios are represented in Figure 1. The ternary plot represents ξ_U, on the red axis, labeled as "Up", ξ_D, on the blue axis, labeled as "Down", and $1 - \xi_D - \xi_U$ on the green axis, labeled as "Stay". Each estimated CUM model can be represented as a point in this plot, with coordinates given by the estimated parameters. The estimated uncertainty is included in the plot as the point size.
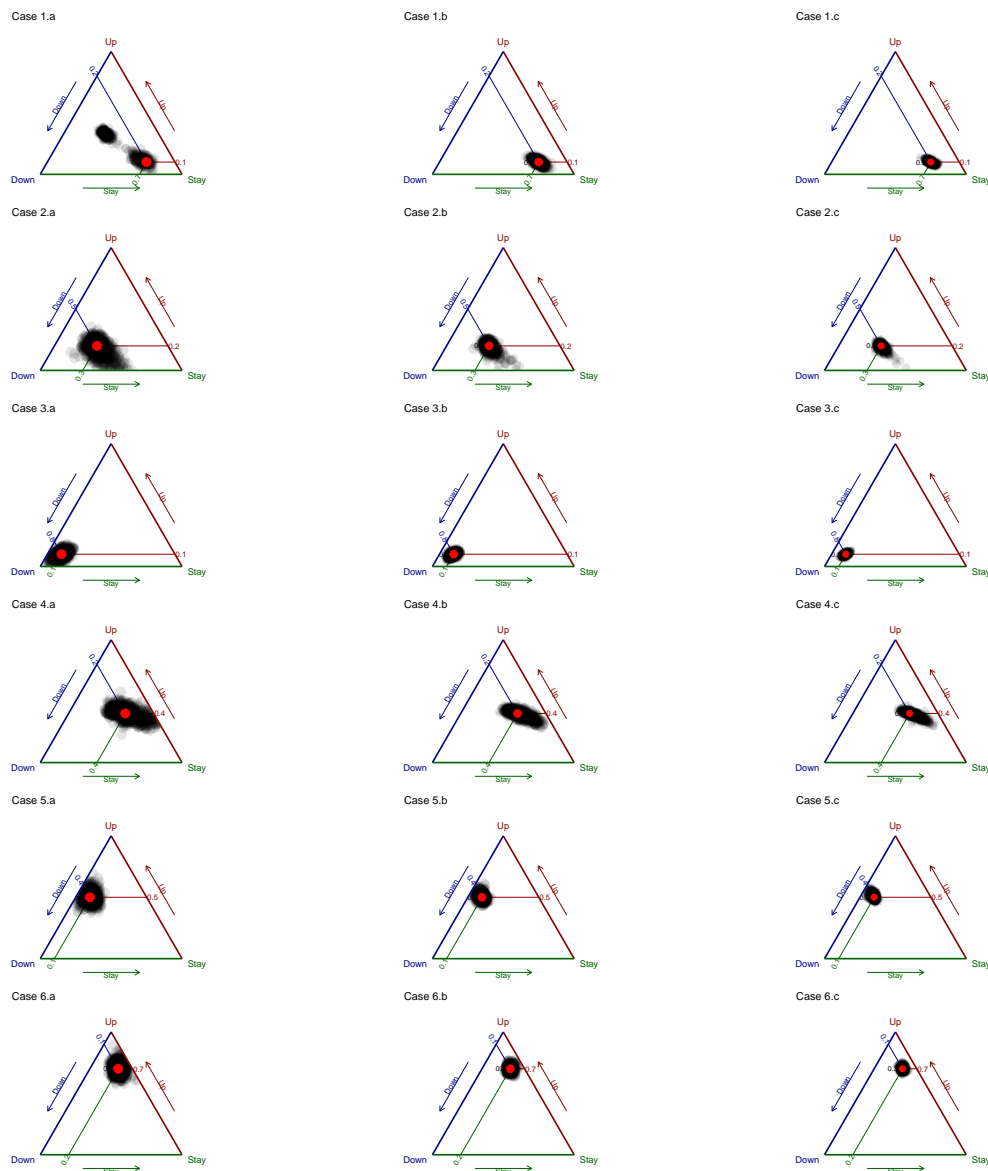

Figure 1: Ternary plots for CUM5 simulations ($n = 1000, iter = 1000$). Red dot: true parameter value; grey dots: estimated values.

In general, the estimated values tend to be quite close to the true parameter value, except for Case 1a, where a portion of the estimated values tend to concentrate on a wrong area of the parameter space. As will be clear in the next Section, this pattern, although still present, tends to be less critical with CUM7. Future research will be addressed to a deep understanding of the source of this anomaly.

Averages and standard errors of the $iter = 1000$ estimated values are reported in Table 3.

The quality of results has been assessed considering three indices: the estimated absolute bias $\widehat{AB}$, the mean squared error $\widehat{MSE}$ and the average *diss* index $\overline{diss}$. $\widehat{AB}$ and $\widehat{MSE}$ are averaged over π, ξ_D, ξ_U. In Table 4 these metrics are shown for all scenarios, while Table 5 shows the best and worst results.

Table 3: Averages and standard errors of the $iter = 1000$ estimated values, CUM5

			a				b				c		
		π	ξ_D	ξ_U	$diss$	π	ξ_D	ξ_U	$diss$	π	ξ_D	ξ_U	$diss$
	True	0.3	0.2	0.1	-	0.5	0.2	0.1	-	0.7	0.2	0.1	-
Case 1	Average	0.6229	0.3418	0.2827	0.0166	0.5035	0.2010	0.1005	0.0082	0.7033	0.2008	0.1005	0.0080
	Std error	0.1664	0.0709	0.0957	0.0115	0.0450	0.0223	0.0224	0.0062	0.0381	0.0170	0.0156	0.0061
	True	0.3	0.5	0.2	-	0.5	0.5	0.2	-	0.7	0.5	0.2	-
Case 2	Average	0.2957	0.4991	0.1864	0.0101	0.4999	0.4982	0.1946	0.0105	0.6972	0.5003	0.1965	0.0100
	Std error	0.0742	0.0490	0.0622	0.0076	0.0752	0.0277	0.0378	0.0079	0.0682	0.0199	0.0261	0.0077
	True	0.3	0.8	0.1	-	0.5	0.8	0.1	-	0.7	0.8	0.1	-
Case 3	Average	0.3004	0.8017	0.0982	0.0076	0.4982	0.8017	0.0997	0.0065	0.7003	0.8009	0.0995	0.0058
	Std error	0.0434	0.0435	0.0304	0.0056	0.0406	0.0253	0.0177	0.0052	0.0333	0.0181	0.0125	0.0044
	True	0.3	0.2	0.4	-	0.5	0.2	0.4	-	0.7	0.2	0.4	-
Case 4	Average	0.3098	0.1952	0.4011	0.0115	0.5031	0.1933	0.3947	0.0110	0.6935	0.1924	0.3952	0.0104
	Std error	0.0832	0.0713	0.0473	0.0088	0.0970	0.0530	0.0339	0.0085	0.0947	0.0398	0.0265	0.0082
	True	0.3	0.4	0.5	-	0.5	0.4	0.5	-	0.7	0.4	0.5	-
Case 5	Average	0.3294	0.3938	0.4963	0.0097	0.5153	0.3991	0.4990	0.0089	0.7023	0.4004	0.5011	0.0089
	Std error	0.0857	0.0393	0.0459	0.0076	0.0966	0.0233	0.0300	0.0068	0.1038	0.0159	0.0233	0.0065
	True	0.3	0.1	0.7	-	0.5	0.1	0.7	-	0.7	0.1	0.7	-
Case 6	Average	0.3004	0.0948	0.7020	0.0083	0.5039	0.1006	0.7001	0.0075	0.6991	0.1001	0.7009	0.0074
	Std error	0.0505	0.0376	0.0458	0.0065	0.0463	0.0217	0.0252	0.0058	0.0396	0.0156	0.0188	0.0053

Table 4: Quality metrics for results of the simulation study, CUM5

	a			b			c		
	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$
Case 1	0.2158	0.0665	0.0166	0.0017	0.0010	0.0082	0.0015	0.0007	0.0080
Case 2	0.0063	0.0040	0.0101	0.0024	0.0026	0.0105	0.0022	0.0019	0.0100
Case 3	0.0013	0.0016	0.0076	0.0013	0.0009	0.0065	0.0006	0.0005	0.0058
Case 4	0.0052	0.0048	0.0115	0.0050	0.0045	0.0110	0.0063	0.0038	0.0104
Case 5	0.0131	0.0040	0.0097	0.0058	0.0037	0.0089	0.0013	0.0039	0.0089
Case 6	0.0025	0.0020	0.0083	0.0015	0.0011	0.0075	0.0007	0.0007	0.0074

Table 5: Best and worst results for CUM5 simulation study

	min			max		
	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$
Case	3c	3c	3c	1a	1a	1a
Value	0.0006	0.0005	0.0058	0.2158	0.0665	0.0166

3.2. CUM7 simulation study

The ternary plots of the CUM7 simulation study for the 18 scenarios are represented in Figure 2. The results of the CUM7 simulation study, assessed with the same metrics used for CUM5
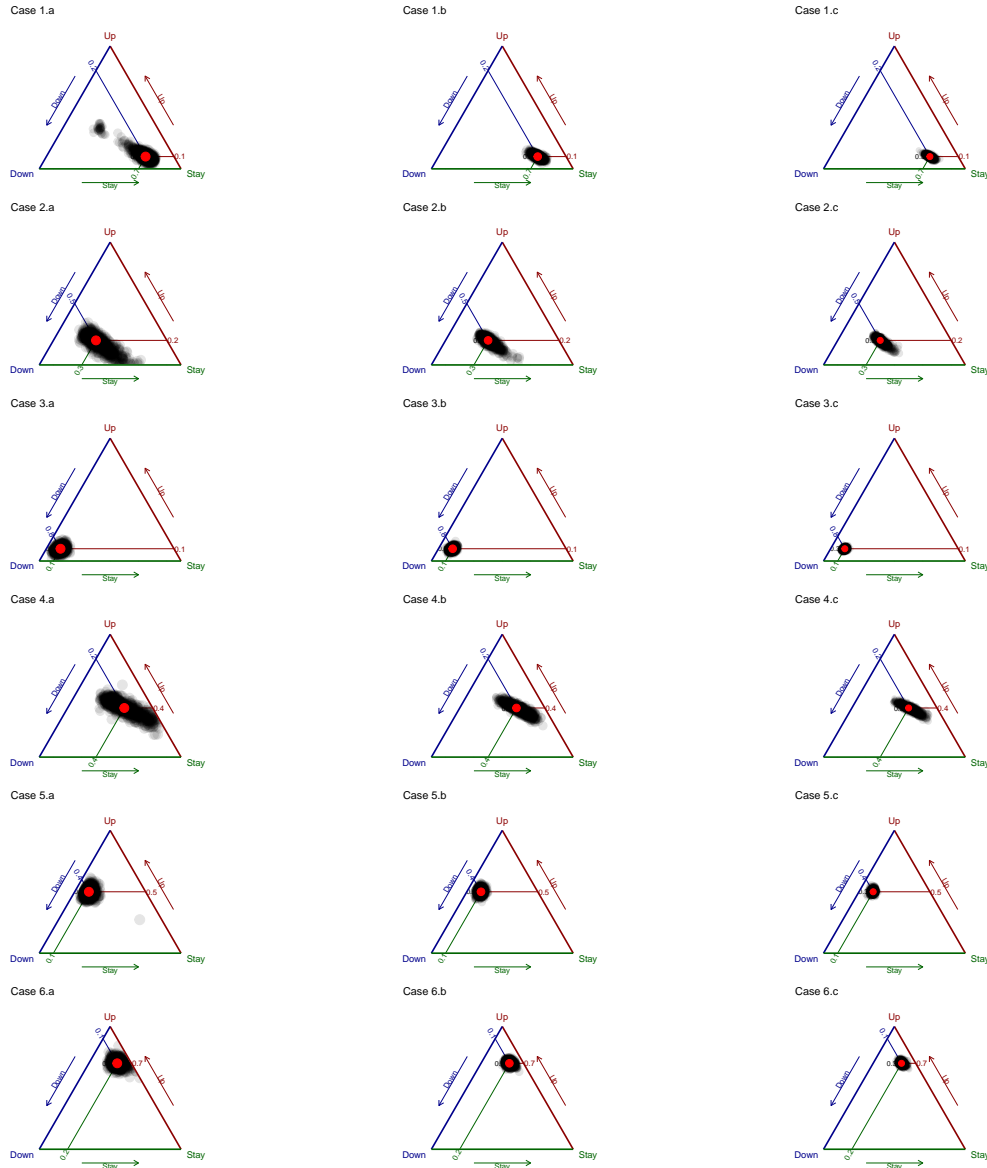

Figure 2: Ternary plots for CUM7 simulations ($n = 1000, iter = 1000$). Red dot: true parameter value; grey dots: estimated values.

(Tables 6, 7 and 8), show an overall slightly better performance, mainly in terms of efficiency. Also with CUM7, Case 1a exhibits a problematic pattern, but, with respect to Figure 1, observations are more concentrated around the true parameter value.

4. Case study

In this section, we present the results obtained from the application of the CUM model to real data obtained by administering a questionnaire to the visitors of the Santa Giulia Museum in Brescia, Italy. The Santa Giulia Museum, included in the UNESCO World Heritage List, is the most important museum in Brescia and unique in Italy and in Europe due to its display concept and location. The data analyzed in this paper were collected during the period April-July 2022 by a survey developed within the activities of the project "Data Science for Brescia (DS4BS) - Arts and cultural places" (`https://bodai.unibs.it/ds4bs/`).

Table 6: Averages and standard errors of the $iter = 1000$ estimated values, CUM7

			a				b				c		
		π	ξ_D	ξ_U	$diss$	π	ξ_D	ξ_U	$diss$	π	ξ_D	ξ_U	$diss$
	True	0.3	0.2	0.1	-	0.5	0.2	0.1	-	0.7	0.2	0.1	-
Case 1	Average	0.3133	0.2105	0.1133	0.0199	0.5025	0.2002	0.1004	0.0185	0.7002	0.2004	0.1000	0.0174
	Std error	0.0519	0.0499	0.0521	0.0097	0.0367	0.0201	0.0188	0.0081	0.0319	0.0145	0.0137	0.0082
	True	0.3	0.5	0.2	-	0.5	0.5	0.2	-	0.7	0.5	0.2	-
Case 2	Average	0.2958	0.4918	0.1822	0.0206	0.4977	0.4966	0.1952	0.0205	0.6972	0.4982	0.1959	0.0205
	Std error	0.0601	0.0504	0.0648	0.0088	0.0556	0.0292	0.0367	0.0090	0.0505	0.0204	0.0274	0.0084
	True	0.3	0.8	0.1	-	0.5	0.8	0.1	-	0.7	0.8	0.1	-
Case 3	Average	0.3016	0.7994	0.0996	0.0172	0.5001	0.8014	0.0991	0.0158	0.6998	0.8003	0.0997	0.0148
	Std error	0.0352	0.0288	0.0230	0.0076	0.0339	0.0183	0.0147	0.0070	0.0291	0.0125	0.0098	0.0067
	True	0.3	0.2	0.4	-	0.5	0.2	0.4	-	0.7	0.2	0.4	-
Case 4	Average	0.3094	0.2016	0.4019	0.0218	0.5000	0.1955	0.3959	0.0211	0.6979	0.1965	0.3976	0.0212
	Std error	0.0685	0.0718	0.0598	0.0094	0.0659	0.0471	0.0403	0.0092	0.0625	0.0348	0.0294	0.0090
	True	0.3	0.4	0.5	-	0.5	0.4	0.5	-	0.7	0.4	0.5	-
Case 5	Average	0.3094	0.3971	0.4956	0.0213	0.5044	0.4007	0.4997	0.0215	0.6971	0.3999	0.5003	0.0211
	Std error	0.0577	0.0331	0.0349	0.0092	0.0619	0.0200	0.0230	0.0090	0.0595	0.0142	0.0155	0.0093
	True	0.3	0.1	0.7	-	0.5	0.1	0.7	-	0.7	0.1	0.7	-
Case 6	Average	0.3018	0.0969	0.6993	0.0180	0.4999	0.0983	0.7009	0.0174	0.6998	0.0987	0.7000	0.0164
	Std error	0.0412	0.0316	0.0307	0.0078	0.0368	0.0184	0.0188	0.0077	0.0313	0.0134	0.0149	0.0073

Table 7: Quality metrics for results of the simulation study, CUM7

	a			b			c		
	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$
Case 1	0.0124	0.0028	0.0199	0.0010	0.0007	0.0185	0.0002	0.0005	0.0174
Case 2	0.0101	0.0036	0.0206	0.0035	0.0018	0.0205	0.0029	0.0012	0.0205
Case 3	0.0009	0.0009	0.0172	0.0008	0.0008	0.0158	0.0003	0.0004	0.0148
Case 4	0.0043	0.0045	0.0218	0.0029	0.0027	0.0211	0.0027	0.0020	0.0212
Case 5	0.0056	0.0019	0.0213	0.0018	0.0016	0.0215	0.0011	0.0013	0.0211
Case 6	0.0019	0.0012	0.0180	0.0009	0.0007	0.0174	0.0005	0.0005	0.0164

Table 8: Best and worst results for CUM7 simulation study

	min			max		
	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$	$\widehat{AB}$	$\widehat{MSE}$	$\overline{diss}$
Case	1c	3c	3c	1a	4a	4a
Value	0.0002	0.0004	0.0148	0.0124	0.0045	0.0218

The dataset contains 665 evaluations expressed by visitors about a question related to the easiness in visiting the museum. The adopted 7-point semantic differential scale ranges from "difficult" to "easy". The absolute frequency distribution of answers is displayed in Figure 3. Results obtained from the application of CUM5 are compared with those obtained from CUM7 and CUB. Specifically, the aim of the case study is twofold: (1) to comparatively assess the results of the CUM and traditional CUB approach and (2) to compare the parameter estimates obtained with CUM5 and CUM7. As for (2), we propose to deal with the original data collected on the 7-point semantic differential scale and then with the same data where some categories are merged using two strategies to create two different 5-point scales. So, data are fitted with CUM7 and CUM5 with the purpose to check the consistency of results.
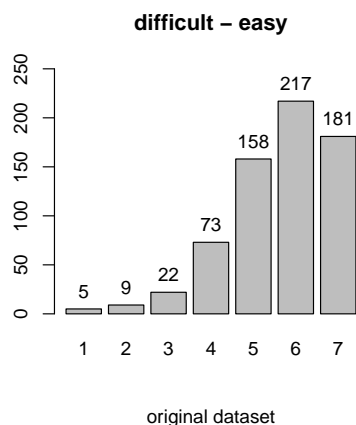
Figure 3: Absolute frequencies of the 7 original ratings for the question about easiness in visiting Santa Giulia museum

In order to adapt the dataset for CUM5 application, categories need to be reduced from seven to five. To this purpose, two different strategies have been implemented:

1. The first one is based on merging ratings two/three and five/six, to maintain the original ending and central categories of the scale. This dataset will be addressed as dataset-1 in the following.

2. The second one is based on merging ratings one/two/three, and leaving unchanged the other categories. This second choice is based on the analysis of frequencies of original ratings (see Figure 3), where categories in the right-hand side of the scale have an appreciably higher frequency than the other categories. This dataset will be addressed as dataset-2 in the following.

The rearranged datasets have the same number of observations as the original one, and the frequency distributions are displayed in Figure 4. The strategies we have implemented to reduce the number of categories from 7 to 5 are arbitrary, even though they are motivated by reasons (1) related to a psychological argument, stating that respondents give much importance to the extreme values of the scale and the middle value, and (2) suggested by data analysis. Other options could also be proposed. Alternatively, we could have considered a dataset obtained from surveys with questions based on 5-point response scales. However, in this case, we would not have been able to fit the CUM7 model.

4.1. Models for the dataset with 7 categories

In this subsection, results of the application of CUB and CUM7 to the original dataset with 7 categories are described and compared. Table 9 reports estimated parameters; the ternary plot chart for CUM7 is displayed in Figure 5. Both the CUM7 and CUB models suggest a low level of uncertainty ($1-\hat{\pi}$ is equal to 0.0509 and 0.0829, respectively). As for feeling, it is quite
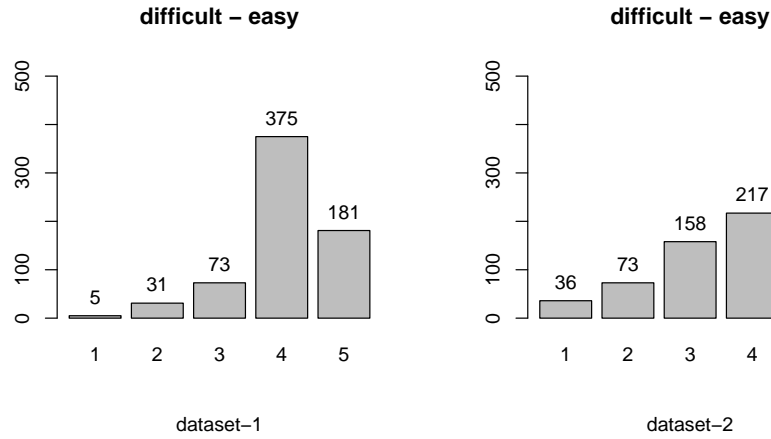
Figure 4: Absolute frequencies of dataset-1 and dataset-2 for the question about easiness in visiting the museum (5 ratings)

high for CUB model ($1 - \hat{\xi} = 0.7929$), while CUM7 recognises the presence of the different components of the assumed DP, namely the probability to move toward "easy" ($\hat{\xi}_U = 0.6527$), to move toward "difficult" ($\hat{\xi}_D = 0.0824$) and to stay still ($1 - \hat{\xi}_U - \hat{\xi}_D = 0.2649$).
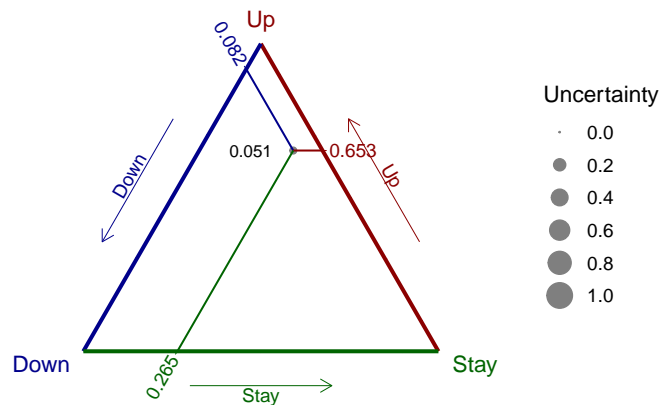

Figure 5: Ternary plot of CUM7 model

Plots of observed versus fitted frequencies for CUM7 and CUB are shown in Figure 6: CUM7 exhibits a better fit than CUB. Diss, BIC and AIC indices for the two models (Table 10) are all lower for CUM7 than for CUB, suggesting that the improved goodness of fit of CUM7 justifies the additional parameter.

4.2. Models for the datasets with 5 categories: dataset-1

In this subsection, CUM5 and CUB are used to fit the ratings of dataset-1. For the CUB approach, a model with shelter on the fourth category was also used, but the shelter parameter turned out to be not significant. Table 11 reports the estimated parameters for CUM5 and CUB models; for CUM5, the ternary plot is in Figure 7. Also with these data, both the
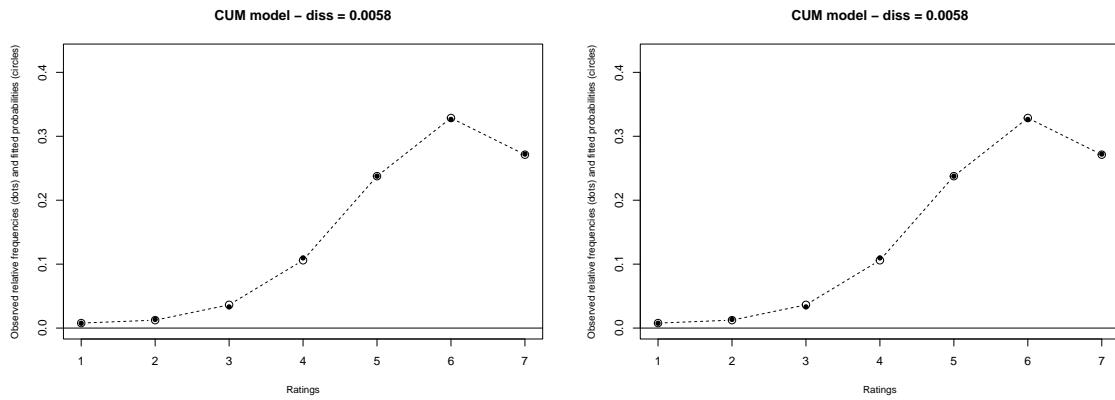
Figure 6: Observed versus fitted frequencies for CUM7 (left) and CUB (right) model

Table 9: Estimated parameters (standard errors in parenthesis) - CUM7 and CUB models fitted to the original dataset with 7 ratings

CUM7			CUB	
π	ξ_D	ξ_U	π	ξ
0.9491	0.0824	0.6527	0.9171	0.2071
(0.0202)	(0.0130)	(0.0140)	(0.0218)	(0.0077)

Table 10: Diss index, BIC and AIC - CUM7 and CUB fitted to the original dataset with 7 ratings

	diss	BIC	AIC
CUM7	0.0058	2030.230	2016.524
CUB	0.0545	2033.283	2024.284

CUM5 and CUB models suggest a low level of uncertainty ($1 - \hat{\pi}$ is equal to 0.0151 and 0.0174, respectively). As for feeling, it is again quite high for CUB model ($1 - \hat{\xi} = 0.7668$), and also in this case CUM5 recognises the presence of the different components of the assumed DP, with estimated values consistent with those obtained with CUM7 fitted to the original dataset ($\hat{\xi}_U = 0.5764$, $\hat{\xi}_D = 0.0245$ and $1 - \hat{\xi}_U - \hat{\xi}_D = 0.3991$).

Plots of observed versus fitted frequencies are displayed in Figure 8. Both CUM5 and CUB are not able to model the large observed frequency in the fourth rating, and this reflects on high values of the diss index, which however is lower for CUM5. Also in this case, according to diss index, BIC and AIC (Table 12), the CUM model outperforms the others.
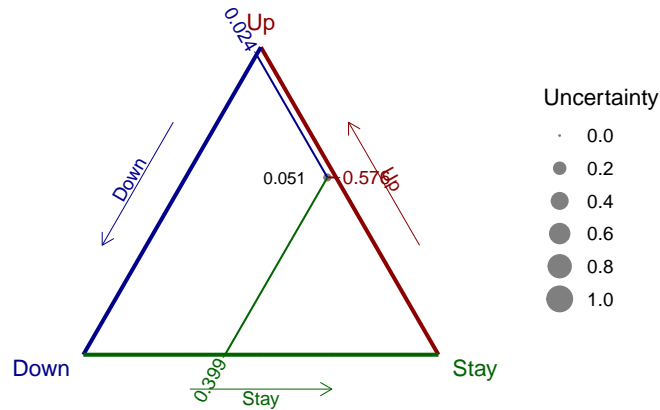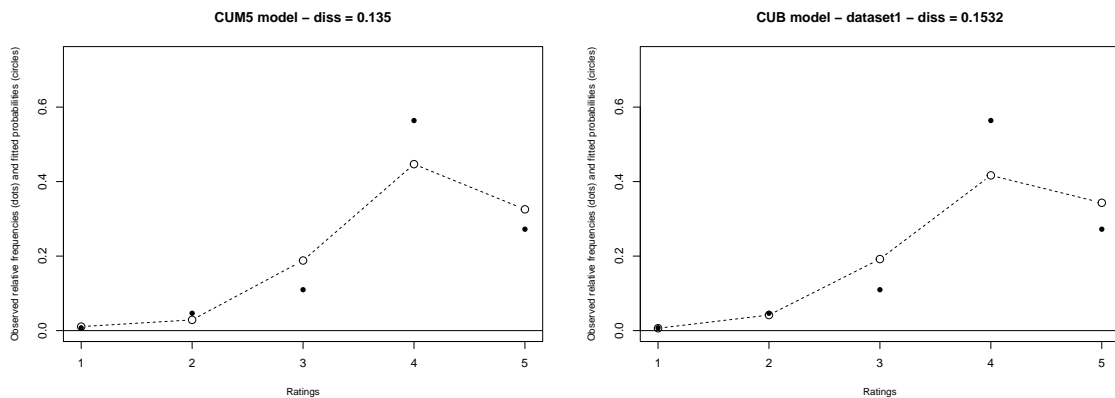

Figure 7: Ternary plot of CUM5 model, dataset-1

Figure 8: Observed vs. fitted frequencies for CUM5 (left) and CUB (right) model, dataset-1

4.3. Models for the datasets with 5 categories: dataset-2

In this subsection, CUM5 and CUB models are fitted to dataset-2. In this case the observed frequencies do not suggest the presence of any shelter effect, so only the basic CUB model is used. The parameter estimates are in Table 13; the ternary plot of CUM is shown in Figure 9. The different aggregation of categories proposed in dataset-2 generates data with higher

Table 11: Estimated parameters (standard errors in parenthesis) - CUM5 and CUB models fitted to dataset-1

CUM5			**CUB**	
π	ξ_D	ξ_U	π	ξ
0.9849	0.0245	0.5764	0.9826	0.2332
(0.0249)	(0.0089)	(0.0150)	(0.0145)	(0.0090)

Table 12: Diss index, BIC and AIC - CUM5 and CUB models fitted to dataset-1

	diss	BIC	AIC
CUM5	0.1350	1539.24	1525.74
CUB	0.1532	1545.10	1536.10

uncertainty, as confirmed by both the CUM5 and CUB approaches ($1 - \hat{\pi}$ is equal to 0.2078 and 0.2874, for CUM5 and CUB, respectively). However, no appreciable difference emerges for feeling, whose parameters have quite similar estimates to those obtained with dataset-1, both for CUB model ($1 - \hat{\xi} = 0.7248$), and for CUM5, except for a sightly higher probability of moving toward "difficult" ($\hat{\xi}_U = 0.5393$, $\hat{\xi}_D = 0.1276$ and $1 - \hat{\xi}_U - \hat{\xi}_D = 0.3331$). The higher value of $\hat{\xi}_D$ can be justified by the fact that the aggregation rule generating dataset-2 (merging of the old categories 1-2-3) moves the middle position of the scale upward. Since the CUM DP assumes that the respondent's reasoning starts from the middle position, if this is moved upward we can reasonably expect a higher probability of moving toward "difficult". So, the different aggregation of categories seems to have modified only the uncertainty assessment, but the feeling measurement remains consistent with that obtained with the other datasets, denoting a very robust assessment of this component.

Figure 10 displays the plots of observed versus fitted frequencies, suggesting a good fit for both the models. According to diss, BIC and AIC indices (Table 14), the CUM model outperforms the other with respect to diss and AIC, while in this case the lowest BIC is reached by CUB.
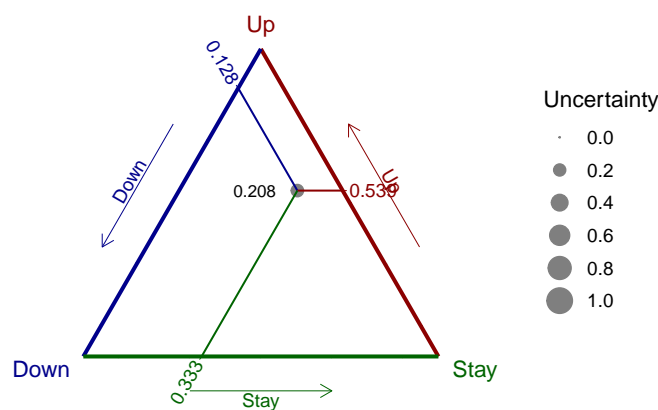

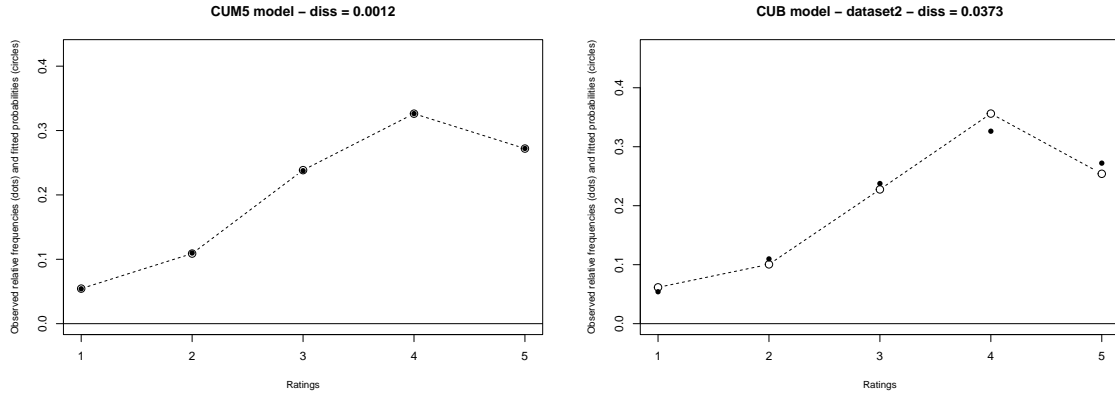Figure 9: Ternary plot of CUM5 model fitted to dataset-2

Figure 10: Observed vs. fitted frequencies for CUM5 (left) and CUB (right) model, dataset-2

Table 13: Estimated parameters (standard errors in parenthesis) - CUM5 and CUB models fitted to dataset-2

CUM5			**CUB**	
π	ξ_D	ξ_U	π	ξ
0.7922	0.1276	0.5393	0.7126	0.2752
(0.0599)	(0.0260)	(0.0201)	(0.0441)	(0.0142)

5. Conclusions

In this paper the CUM model, conceived in the framework of the CUB class, is recalled with reference to its original formulation (suited for multi-point semantic differential with 7 categories, CUM7) and then extended to the case of 5 categories (CUM5). R functions implementing this new proposal have been developed and made available on the website `https://bodai.unibs.it/cub/`.

The results from a simulation considering the application of CUM5 to 18 scenarios based on different parameter values are reported and compared to the results from an equivalent CUM7 simulation, showing how fitting measures are similar in the two cases, with a single case that requires further investigation.

A case study concerned with the the evaluation of the visitor experience at the Santa Giulia Museum is proposed. The original dataset with 7 ratings was analysed via CUM7 and CUB. Then, the original dataset has been transformed into a 5-point scale, following two different strategies, and analysed by means of CUM5 and CUB. The idea was to compare the CUM to the CUB approach and also to check the robustness of results with respect to different aggregations of the original dataset with 7 categories into two datasets with 5 categories.

From the point of view of the comparison of the different approaches, in general the CUM model outperforms CUB, apart from a single case where CUB exhibits a lower BIC value. So, the additional parameter of CUM with respect to CUB seems to provide significant information and goodness-of-fit improvement.

As for parameter estimates, in the examined case study, the different aggregation rule of the response categories impacts on the uncertainty measurement, while the assessment of the

Table 14: Diss index, BIC and AIC - CUM5 and CUB models fitted to dataset-2

	diss	BIC	AIC
CUM5	0.0012	1963.29	1949.79
CUB	0.0373	1960.79	1951.79

feeling component remains stable, except for a small, easily interpretable, difference. So, both CUM and CUB exhibit high robustness with respect to different manipulations of data.

Next steps will be concerned with the deeper investigation of the single simulation case producing estimates far from actual parameters values both for CUM5 and, to a lesser degree, for CUM7. In particular, this raises concerns about the identifiability of the CUM5 model over the entire parameter space, posing an unresolved issue that necessitates further investigation. While awaiting conclusive results, we recommend considering CUM7 whenever feasible, as it appears to be unaffected by identifiability concerns. Additionally, we propose estimating the model by first verifying the uniqueness of estimates using different starting values for the EM algorithm, as done in this paper.

Acknowledgements

References

Agresti A (2010). *Analysis of Ordinal Categorical Data*, volume 656. John Wiley & Sons.

Agresti A (2013). *Categorical Data Analysis.* 3rd edition. Wiley, New York.

Biasetton N, Disegna M, Barzizza E, Salmaso L (2023). "A New Adaptive Membership Function with CUB Uncertainty with Application to Cluster Analysis of Likert-type Data." *Expert Systems with Applications*, **213**, 118893. doi:https://doi.org/10.1016/j.eswa.2022.118893.

Cappelli C, Simone R, Iorio FD (2019). "CUBREMOT: A Tool for Building Model-Based Trees for Ordinal Responses." *Expert Systems with Applications*, **124**, 39–49. doi:https://doi.org/10.1016/j.eswa.2019.01.009.

Cerulli G, Simone R, Di Iorio F, Piccolo D, Baum CF (2022). "Fitting Mixture Models for Feeling and Uncertainty for Rating Data Analysis." *The Stata Journal*, **22**(1), 195–223. doi:https://doi.org/10.1177/1536867X221083927.

Dawis RV (1987). "Scale Construction." *Journal of Counseling Psychology*, **34**(4), 481. doi:https://doi.org/10.1037/0022-0167.34.4.481.

D'Elia A, Piccolo D (2005). "A Mixture Model for Preference Data Analysis." *Computational Statistics & Data Analysis*, **49**, 917–934. doi:https://doi.org/10.1016/j.csda.2004.06.012.

Manisera M, Zuccolotto P (2014a). "Modeling Rating Data with Nonlinear CUB Models." *Computational Statistics & Data Analysis*, **78**, 100–118. doi:https://doi.org/10.1016/j.csda.2014.04.001.

Manisera M, Zuccolotto P (2014b). "Nonlinear CUB Models: The R Code." *Statistics & Applications*, **12**(2), 205–223.

Manisera M, Zuccolotto P (2015). "On the Identifiability of Nonlinear CUB Models." *Journal of Multivariate Analysis*, **140**, 302–316. doi:https://doi.org/10.1016/j.jmva.2015.05.011.

Manisera M, Zuccolotto P (2017). "Estimation of Nonlinear CUB Models Via Numerical Optimization and EM Algorithm." *Communications in Statistics-Simulation and Computation*, **46**(7), 5723–5739. doi:https://doi.org/10.1080/03610918.2016.1175622.

Manisera M, Zuccolotto P (2019). "Discussion of: The Class of CUB Models: Statistical Foundations, Inferential Issues and Empirical Evidence, by Domenico Piccolo And Rosaria Simone." *Statistical Methods & Applications*, **28**, 465–470. doi:https://doi.org/10.1007/s10260-019-00465-x.

Manisera M, Zuccolotto P (2022). "A Mixture Model for Ordinal Variables Measured on Semantic Differential Scales." *Econometrics and Statistics*, **22**, 98–123. ISSN 2452-3062. doi:https://doi.org/10.1016/j.ecosta.2021.07.002. The 2nd Special issue on Mixture Models.

Nelder JA, Mead R (1965). "A Simplex Algorithm for Function Minimization." *Computer Journal*, **7**, 308–313. doi:https://doi.org/10.1093/comjnl/7.4.308.

Piccolo D (2003). "On The Moments of a Mixture of Uniform and Shifted Binomial Random Variables." *Quaderni di Statistica*, **5**, 85–104.

Piccolo D, Simone R (2019). "The Class of CUB Models: Statistical Foundations, Inferential Issues and Empirical Evidence." *Statistical Methods & Applications*, **28**, 389–493. doi:https://doi.org/10.1007/s10260-019-00461-1. (with discussion and rejoinder).

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Simone R (2021). "An Accelerated EM Algorithm for Mixture Models with Uncertainty for Rating Data." *Computational Statistics*, **36**(1), 691–714. doi:https://doi.org/10.1007/s00180-020-01004-z.

Simone R, Cappelli C, Di Iorio F (2019). "Modelling Marginal Ranking Distributions: The Uncertainty Tree." *Pattern Recognition Letters*, **125**, 278–288. doi:https://doi.org/10.1016/j.patrec.2019.04.026.

Simone R, Tutz G, Iannario M (2020). "Subjective Heterogeneity in Response Attitude for Multivariate Ordinal Outcomes." *Econometrics and Statistics*, **14**, 145–158. doi:https://doi.org/10.1016/j.ecosta.2019.04.002.

Affiliation:

Marica Manisera
Department of Economics and Management
Big&Open Data Innovation Laboratory (BODaI-Lab)
University of Brescia
25125 Brescia, Italy
E-mail: marica.manisera@unibs.it
URL: https://www.unibs.it/en/ugov/person/3750