# A dissimilarity-based splitting criterion for CUBREMOT

Carmela Cappelli, Rosaria Simone and Francesca Di Iorio

**Abstract** CUBREMOT (CUB REgression MOdel Trees) is a model-based approach to grow trees for ordinal responses that relies on a class of mixture models for evaluations and preferences (CUB). The original proposal considers deviances in log-likelihood to partition observations. In the present paper a new splitting criterion is introduced that, among the significant splitting variables, chooses the one that maximizes a dissimilarity measure. This choice is tailored to generating child nodes as far apart as possible with respect to the estimated probability distributions. An application to real data on Italians' trust towards the European Parliament taken from the official survey on daily life conducted by the Italian National Institute of Statistics (ISTAT) in 2015 is presented and discussed in comparison with alternative methods.
**Abstract** *Nel presente lavoro viene proposto un nuovo criterio di split per la procedura* CUBREMOT *(CUB REgression MOdel Trees).* CUBREMOT *è uno strumento per crescere alberi per risposte ordinali ai cui nodi sono associati modelli mistura per le valutazioni e preferenze (modelli* CUB*) e che utilizza un criterio di split basato sulla differenza in log-verosimiglianza. Il criterio di split alternativo che viene qui introdotto utilizza invece un indice di dissimilarità per generare, attraverso lo split di un nodo padre, nodi figli che siano il più distanti possibile in termini della distribuzione di probabilità stimata. La validità dell'approccio e il confronto con altri metodi sono mostrati mediante l'applicazione a dati reali sulla fiducia verso il Parlamento Europeo sulla base dell'indagine multiscopo condotta nel 2015 dall'ISTAT.*

**Key words:** Tree -based methods; Ordinal Responses; Dissimilarity measure

Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò, 22, 80138 Napoli, e-mail: carcappe@unina.it, e-mail: rosaria.simone@unina.it, e-mail: fdiiorio@unina.it

# 1 Introduction

In the spirit of the model-based partitioning approach [9], CUBREMOT [2, 3] is a tool for growing trees for ordinal responses in which every node is associated with a CUB model [4]. This approach to model preferences, judgements and perceptions is based on the idea that discrete choices arise from a psychological process that involves a personal *feeling* and an inherent *uncertainty* both possibly related to explanatory covariates.

The splitting criterion employed in CUBREMOT computes the log-likelihood increment from the father node to the child nodes for each possible split, and at the given step chooses the one that maximizes such deviance. Thus, this criterion selects the covariate that entails the most plausible values for CUB parameters in the child nodes among the variables that are significant for at least one of the model components at the father node.

We propose a further splitting criterion that focuses on the dissimilarity between child nodes, aiming at generating child nodes as far apart as possible with respect to the probability distributions estimated by CUB models. Both splitting criteria generate a model-based tree whose terminal nodes provide different profiles of respondents, which are classified into nodes according to levels of feeling and uncertainty conditional to the splitting covariates. In what follows, we briefly recall the main features of CUBREMOT, then we present the new splitting criterion and we illustrate the results of an application to data from the official survey on daily life conducted by the Italian National Institute of Statistics (ISTAT) in 2015 focusing on Italians' trust towards the European Parliament.

# 2 Background and Methodology

CUB models paradigm [4] designs the data generating process yielding to a discrete choice on a rating scale as the combination of a *feeling* and an *uncertainty* component. The resulting mixture prescribes a shifted Binomial distribution for feeling to account for substantial likes and agreement and assigns a discrete Uniform distribution for uncertainty to shape heterogeneity. Then, if $R_i$ denotes the response of the $i$-th subject to a given item of a questionnaire,

$$Pr(R_i = r|\pi_i, \xi_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r}(1-\xi_i)^{r-1} + (1-\pi_i)\frac{1}{m}, \quad r = 1, \dots, m,$$

where the model parameters $\pi_i$ and $\xi_i$ are called uncertainty and feeling parameter, respectively. Covariates may be included in the model in order to relate feeling and/or uncertainty to respondents' profiles. Customarily, a logit link is considered:

$$logit\,(\pi_i) = \boldsymbol{x}_i\boldsymbol{\beta}; \qquad logit\,(\xi_i) = \boldsymbol{w}_i\boldsymbol{\gamma}\,, \tag{1}$$

where $\boldsymbol{x}_i, \boldsymbol{w}_i$ are the values of selected explanatory variables for the $i$-th subject. If no covariate is considered neither for feeling nor for uncertainty, then $\pi_i = \pi$ and $\xi_i = \xi$ are constant among subjects. Estimation of CUB models relies on likelihood methods and on the implementation of the Expectation-Maximization (EM) algorithm.
In CUBREMOT, CUB models are employed in the top-down partitioning algorithm that grows the tree as follows. According to binary recursive partitioning, each of the available covariates is sequentially transformed into suitably splitting variables or binary questions which are Boolean condition on the value (or categories) of the covariate where the condition is either satisfied ("yes") or not satisfied ("no") by the observed value of that covariate (for details see [1] ). In this respect any split $s$ can be seen as a dummy variable.
Then, for a given node $k \geq 1$ with size $n_k$, a CUB without covariates is fitted, whose log-likelihood at the final ML estimates $(\hat{\pi}_k, \hat{\xi}_k)$ is denoted by $\mathscr{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k)$. Then, a CUB with splitting variable $s$ is tested: if it is significant for at least one component, it implies a split into a left and right child nodes that will be associated with the conditional distributions $R|s = 0$ with parameter values $(\hat{\pi}_{2k}, \hat{\xi}_{2k})$ and $R|s = 1$ with parameter values $(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})$, respectively, Thus, the splitting criterion proposed in [2, 3] at the given step chooses the split that maximizes the deviance:

$$\Delta\mathscr{L}_k = \left[\mathscr{L}_{n_{2k}}(\hat{\pi}_{2k}, \hat{\xi}_{2k}) + \mathscr{L}_{n_{2k+1}}(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})\right] - \mathscr{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k). \tag{2}$$

Indeed, such difference measures the improvement in log-likelihood yielded by the inclusion of the significant splitting variable and the best split, being associated with the maximum log-likelihood increment, provides the child nodes characterized by the most plausible values for CUB parameters.

Here we propose an alternative splitting criterion based on the concept of dissimilarity between child nodes: the proposal considers a proper version of the normalized index proposed by [8] that compares the estimated probability distribution with the observed relative frequencies and it is generally considered in the framework of CUB models as a goodness of fit measure. Specifically, aiming at the generation of child nodes that are the farthest apart from each other in terms of distribution of the responses, in the set $\mathscr{S}_k = \{s_{k,1}, \dots, s_{k,l}\}$ of the $l$ significant splitting variables for node $k$, a CUBREMOT is grown by choosing, at each step, the split maximizing the distance between the estimated CUB probability distributions $\hat{p}_{2k}$ and $\hat{p}_{2k+1}$ for the child nodes in terms of the dissimilarity measure:

$$Diss(2k, 2k+1) = \frac{1}{2} \sum_{r=1}^{m} |\hat{p}_{2k} - \hat{p}_{2k+1}|. \tag{3}$$

The choice of this normalized index entails that, as long as CUB models estimated at the child nodes provide an adequate fitting, the splitting variable generates an optimal partition of the father node in terms of the chosen distance. In particular, the resulting terminal nodes determine well-separated profiles of respondents, in terms of both feeling (agreement, preferences, and so on) and uncertainty (indecision, het-

erogeneity).
Note that, up to now, no retrospective pruning is implemented for CUBREMOT , as two natural stopping rules are available: node partitioning stops (i.e. a node is declared terminal) if either none of the available covariates is significant or the sample size is too small to support a CUB model fit.

## 3 Application

In order to grow a CUBREMOT using the defining splitting criterion in [2, 3], data from the yearly multiscope survey on daily life run by ISTAT in 2015 have been considered. The data set and its detailed description are available at:
www.istat.it/it/archivio/129916.
Here, the chosen response variable is *Trust in EU Parliament* ($TEP$ for short) and it has been collected on a Likert type scale with 11 categories, ranging from 0 = 'I totally distrust it', to 10 = "I have absolute trust in it": as customarily, it has been forward shifted to the range 1-11 for CUB models fitting. For illustrative purposes, only a subset of the available covariates has been given in input to the procedure. Moreover, for the sake of saving space, the CUBREMOT growth has been stopped to three levels and only node 7 and 10 have been declared terminal according to the stopping rules defined in section 2. The tree is displayed in Figure 1, highlighting that the following covariates affect evaluations and discriminate response patterns:

1. *Political Talk* ($PT$), an ordinal factor with levels from 1= "On daily basis" to 6 = "Never" to assess the frequency of involvement in political talks and discussion;
2. *Economic Satisfaction* (*ES*): an ordinal factor asking interviewees to assess their satisfaction towards their wealth status within the previous 12 months, on a balanced scale with levels 1 ="Very satisfied", 2 = "Fairly Satisfied", 3 = "Little Satisfied" up to 4 = "Not at all satisfied";
3. *General Satisfaction* (*GS*): an ordinal factor asking interviewees to assess their overall life satisfaction on a rating scale ranging from levels 0 ="Not all satisfied" up to 10 = "Extremely Satisfied";
4. *Trust in Italian Parliament* ($TIP$): an ordinal variable asking respondents to rate their perceived trust in the Italian Parliament, collected on the same scale as the chosen response variable $TEP$.

For each node, the number of observations $n_k$, the estimated CUB parameters $\pi$ and $\xi$, as well as the dissimilarity between the estimated CUB probabilities for the descending split (*DissB*) are reported while Figure 2 shows fitted (vertical bars) and estimated distributions at selected nodes by reporting also their dissimilarity. We might conclude that *Trust towards the Italian Parliament* plays a prominent role in the understanding and modelling of *Trust towards the European Parliament* but it interacts with the perception of economic well-being and general satisfaction as well as with the direct involvement of the respondents in political talks. In addition, having chosen the CUB paradigm as the root of the model-based approach, nodes

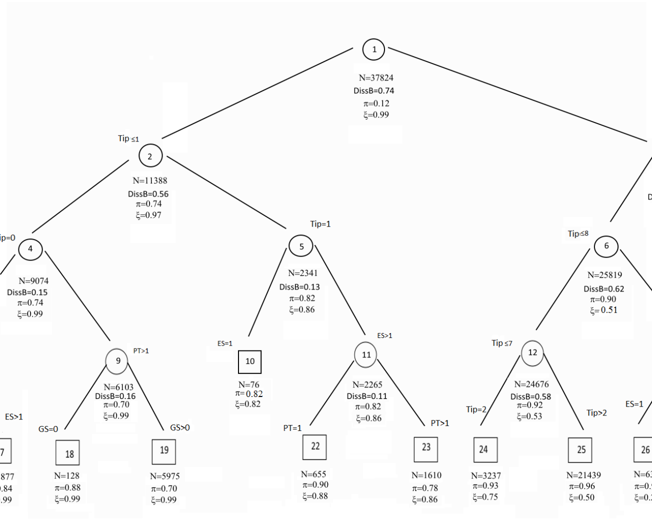

**Fig. 1** CUBREMOT for *Trust for European Parliament* assuming a dissimilarity splitting rule

can be discriminated both in terms of trust (feeling) and in terms of indecision and heterogeneity of the respondents. For instance, at the first split the procedure recognizes a major difference between those with an extremely low trust towards the Italian Parliament ($TIP \leq 1$) and those giving higher evaluations: as shown by Figure 2, these two groups correspond to people with extremely low trust towards the EU Parliament and people with intermediate evaluations, respectively. From Nodes 10, 13, one derives that satisfaction for the economic status is associated, in general, with a higher trust (as measured by $1-\xi$) but also with a higher indecision (Node 16). People claiming to be always involved in political talks ($PT = 1$) are in general more resolute and homogeneous in the responses (Node 8 against Node 9) and also less trustful of the European Parliament (Node 22 against Node 23).

## 4 Final remarks

In a comparative perspective, a tree for the chosen response variable has been grown using the `RpartScore` package [5], which implements the work of [7] to deal with ordinal responses as an extension of the `Rpart` package. In this respect, since it is a common belief that, when the number of categories is high as in the selected case study, the response can be treated as numeric, a tree using standard `Rpart` has also been grown. In both cases, the only splitting variable able to grow the tree is $TIP$, no other covariate is selected even when relaxing the parameters that control the growth of the tree. On the contrary, the proposed approach allows to disclose several drivers of the responses at different levels and with different strength. Also notice that the

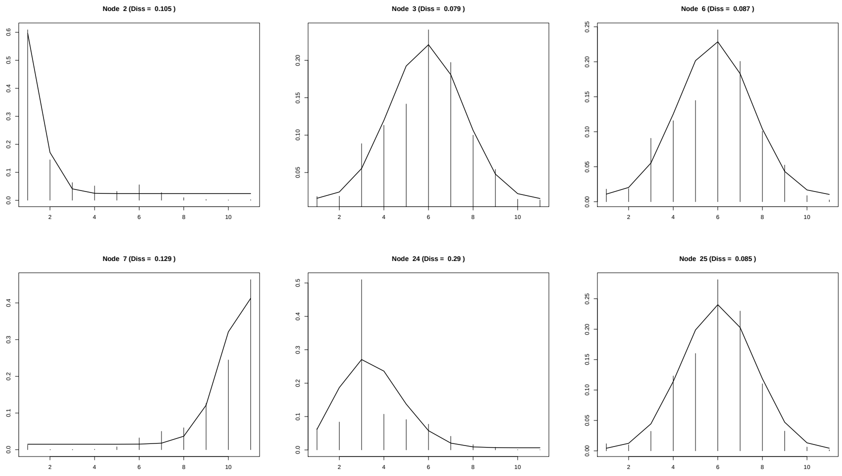

**Fig. 2** Observed and fitted probability distributions at selected nodes

dissimilarity-based splitting criterion grows a different tree with respect to the log-likelihood splitting criterion but they both allow to disentangle various determinants of the response assuming specific decision rules on variable importance. Ongoing research involves a deep comparison with other tree-based methods based on simulation studies as well as the implementation of retrospective pruning while future research will be devoted to a more flexible modelling of the node distributions by considering extensions of CUB models.

## References

1. Breiman L., Friedman J.H., Olshen R.A, Stone C.J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks: Monterey (CA).
2. Cappelli, C., Simone, R. and Di Iorio, F. Model-based trees to classify perception and uncertainty: analyzing trust in European institutions, *under review*, 2017.
3. Cappelli, C., Simone, R., Di Iorio, F. (2017). Growing happiness: a model-based tree. In SIS 2017. Statistics and Data Science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society, Florence 28–30 June 2017, 261-266.
4. D'Elia A., Piccolo D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**, 917–934.
5. Galimberti G., Soffritti G., Di Maso M. (2012). Classification Trees for Ordinal Responses in R: The RpartScore Package, *Journal of Statistical Software*, **47**, 1–25.
6. Iannario M., Piccolo D., Simone, R. (2017). CUB: A Class of Mixture Models for Ordinal Data. R package, version 1.1.1 http://CRAN.R-project.org/package=CUB.
7. Picarretta R. (2008). Classification trees for ordinal variables, *Computational Statistics*, **23**, 407–427.
8. Leti, G. (1983). *Statistica descrittiva*. Il Mulino, Bologna.
9. Zeileis A., Hothorn T., Hornik K. (2008). Model-Based Recursive Partitioning, *Journal of Computational and Graphical Statistics*, **17**, 492–514.