An overview of tree-based methods for ordinal data

Rosaria SIMONE

Università degli Studi di Napoli Federico II

May 25th, 2023

Università degli Studi di Brescia
Workshop on Ordinal Data

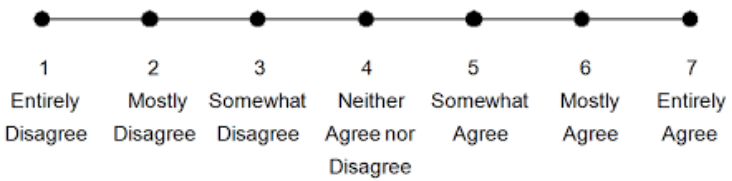
Outline

Ordinal data from surveys

Rating:
Responses convey the `level` of a "perception"

risk, taste, fear, agreement, . . .

1. *How likely is it that you would recommend this brand to a friend or colleague?* (Net Promoter Score)
2. *Does your family easily make ends meet?*
3. *How do you feel safe in the place where you live?*
4. . . .



Ranking:
Responses convey the `location/preference` of the "object" in a given ordered list:

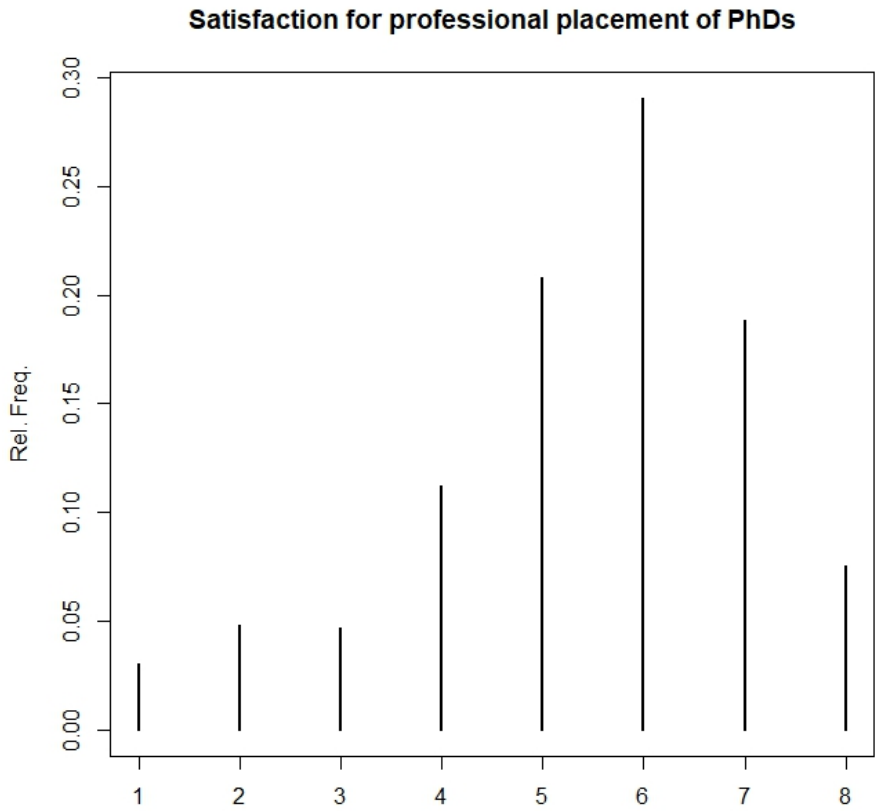
items, products, sports, applicants, sentences, teams, songs, . . .

"In your opinion, what is the probability that you will reach age 75?" Please provide a value between 0 (impossible event) and 100 (certain event).

R	Subjective survival probability	Ordinal interpretation
1	$0.00 \leq Pr(S) \leq 0.05$	`IMPOSSIBLE/Almost IMPOSSIBLE`
2	$0.05 < Pr(S) \leq 0.25$	`LOW`
3	$0.25 < Pr(S) \leq 0.45$	`Moderately LOW`
4	$0.45 < Pr(S) \leq 0.55$	`About FIFTY/FIFTY`
5	$0.55 < Pr(S) \leq 0.75$	`Moderately HIGH`
6	$0.75 < Pr(S) \leq 0.95$	`HIGH`
7	$0.95 < Pr(S) \leq 1.00$	`SURE/Almost SURE`

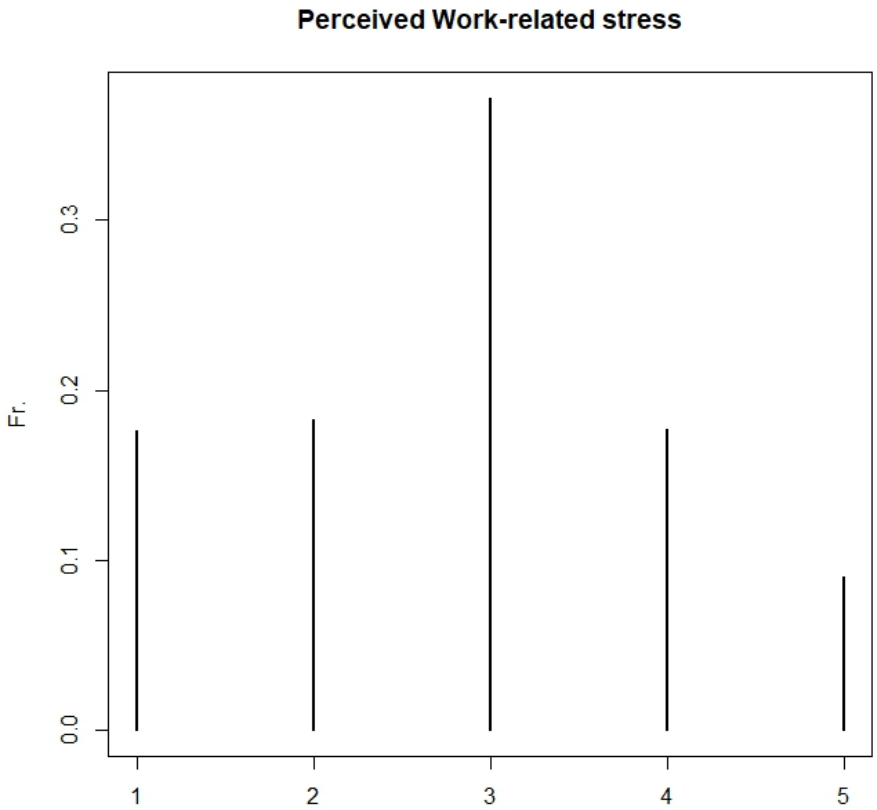
Data example 1: Satisfaction for professional placement of Italian PhDs

- Consider the overall satisfaction for the professional placement of Italian PhDs of cohorts 2012 and 2014, collected within the survey run by the Italian National Office (ISTAT) to investigate the professional placement of PhD (available at `https://www.istat.it/it/archivio/87536`).
- All the ratings were collected on a scale with 10 ordered categories: the rating scale has been subsequently modified to a scale with 8 categories due to zero-scores observed in certain categories, so that higher scores along the response scale corresponds to higher levels of satisfaction.



Data example 2: Work-related stress

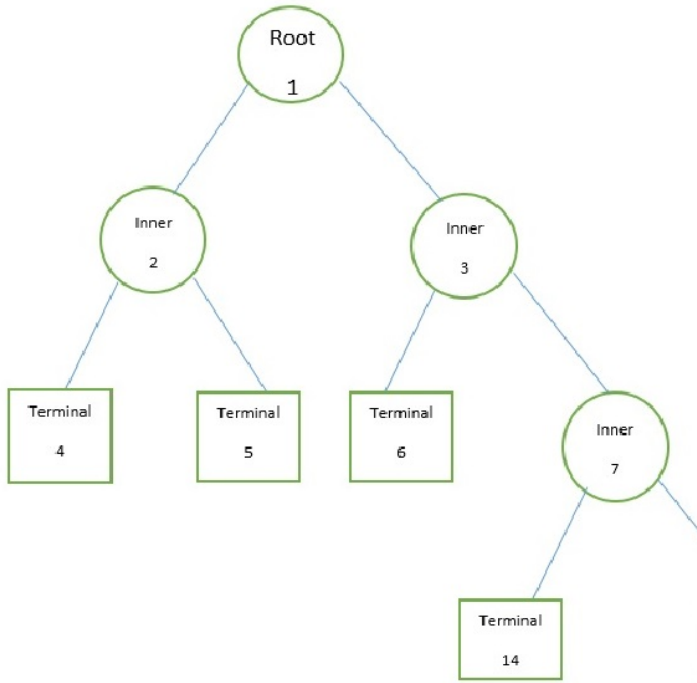
- Data from the 5th European Working Condition Survey carried out by Eurofound in 2010 on working conditions for the EU28.
- Consider $n = 972$ responses for Italy to the question 'Do you experience stress in your work?' measured on a $m = 5$ wording-type scale: 'Always', 'Most of the time', 'Sometimes', 'Rarely', 'Never', coded from 1 to 5.



Tree methods

Top-down greedy recursive binary partitioning

- Top-down: from the whole set of observations to smaller sets of observations, according to a *hierarchy*
- Greedy: The optimality of the splitting criterion is considered at local level only (at each node), and not with reference to subsequent steps and the final tree
- Recursive: The procedure is applied starting from the root node, and then for each descendant, following the same methods and criteria, until a *stopping criterion* is eventually met
- Binary partitioning: Each node (set of observations) is partitioned into two subsets, on the basis of certain values of a predictor (splitting variables)



From predictors to splitting variables

If X is a predictor, *candidate splitting variables* are created at each step:

- quantitative (numeric): then build a splitting variable according to the rule $X \leq s$ or $X > s$, for each $s \in Range(X)$:
  1. X = age: $X \leq 18$ or $X > 18$, $X > 30$ or $X < 30$, ...
  2. X number of purchases: $X \leq 5$ or $X > 5$, ...
- categorical ordinal, with m levels: then build a splitting variable according to the rule $X \leq c_j$ or $X > c_j$, for all ordered level c_j: $c_1 \prec c_2 \prec \cdots \prec c_m$ (for a total of $m-1$ splitting binary variables):
  1. X = customer satisfaction on $m = 5$ rating scales (1 ='very unsatisfied', 2 ='unsatisfied', 3 ='indifferent', 4 = 'satisfied', 5 ='very satisfied'): X very unsatisfied or unsatisfied ($X \leq 2$), or X indifferent or satisfied ($X \geq 3$), ...
- categorical nominal, with m levels $c_1, \ldots, c_m$: then build a splitting variable according to the rule $X \in S$ or $X \notin S$, for all non-trivial subset S of $\{c_1, \ldots, c_m\}$, for a total of $2^{m-1} - 1$ splitting binary variables:
  1. X = gender (two levels: M, F), only one splitting variable $X = M$ or $X = F$
  2. X = marital status (4 levels: single, in a relationship, divorced, widower): 7 candidate splitting variables:
     1. X = Single VS $X \neq$ Single
     2. X = In a relationship VS $X \neq$ in a relationship
     3. X = Divorced VS $X \neq$ Divorced
     4. X = Widower VS $X \neq$ Widower
     5. X = Single or in a relationship VS X = divorced or widower
     6. X = Single or divorced VS X = in a relationship or widower
     7. X = Single or widower VS X = in a relationship or divorced

CART algorithm (Breiman, Friedman, Olshen & Stone (1984))

CART: Classification And Regression Trees[1]:

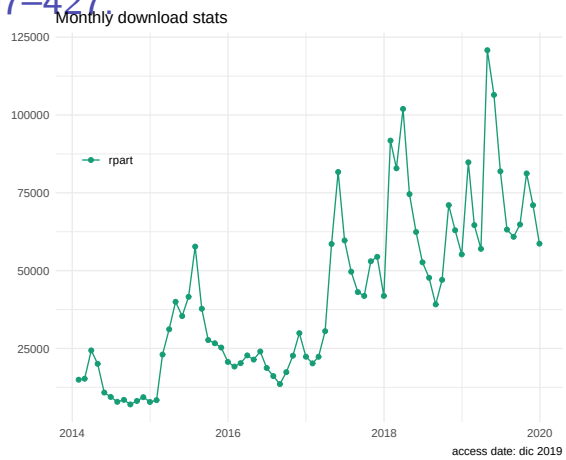
- Classification: for qualitative responses; Regression: for numerical responses

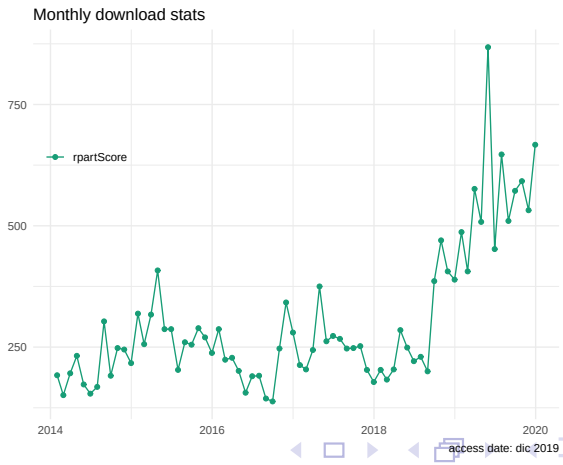
Breiman, L., Friedman J.H., Olshen,R.A. and C.J Stone. (1984). Classification and Regression Trees. Chapman & Hall CRC, Boca Raton

Therneau T. and B. Atkinson (2018). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. https://CRAN.R-project.org/package=rpart

Galimberti G., Soffritti G., Di Maso M. (2012). Classification Trees for Ordinal Responses in R: The rpartScore Package. *Journal of Statistical Software*, **47**(10), 1-25. URL http://www.jstatsoft.org/v47/i10/.

Picarretta, R. (2008). Classification trees for ordinal variables. *Computational Statistics*, **23**(3): 407–427.





Prediction and error for nodes

Let $l_1, \ldots, l_T$ be the terminal nodes (leaves) of a tree.

Node	l_1	l_2	$\cdots$	$\cdots$	l_T
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Prediction	$\hat{y}^{(l_1)}$	$\hat{y}^{(l_2)}$	$\ldots$	$\ldots$	$\hat{y}^{(l_T)}$
Error	$e^{(l_1)}$	$e^{(l_2)}$	$\ldots$	$\ldots$	$e^{(l_T)}$

The prediction for a new observation corresponding to leaf l (in terms of covariate values learnt from tree branches down to l) is:

$$\hat{y}^{(l)} = \begin{cases} Mode^{(l)} = \underset{j=1,\ldots,m}{argmax}\{f_1^{(l)}, \ldots, f_m^{(l)}\}, & \text{for classification trees} \\ \bar{y}^{(l)} = \dfrac{1}{|l|} \sum_{i \in l} y_i, & \text{for regression trees} \end{cases}$$

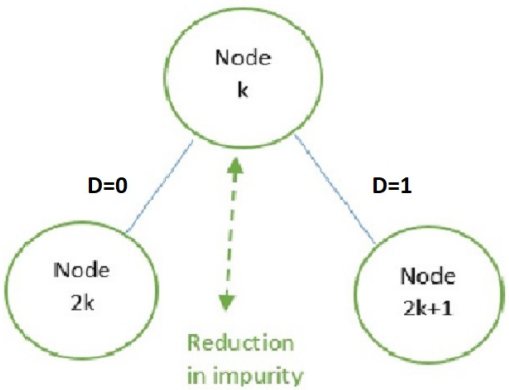
Splitting rule and Generalized Gini impurity function

At node k, the procedure chooses the splitting variable among the candidate ones so that the decrement in impurity is maximized (namely, chose the split that generate more homogeneous children nodes):

$$\Delta i(k) = i(k) - \left(\frac{n_{2k}}{n_k} i(2k) \; + \; \frac{n_{2k+1}}{n_k} i(2k+1) \right)$$

where:

$$i(k) = \begin{cases} RSS_k = \sum_{h=1}^{n_k} \left(y_h - \bar{y}^{(k)} \right)^2, & \text{for regression trees} \\ \mathcal{G}_k^{\star} = \sum_{j=1}^{m} \sum_{h=1}^{m} L(j,h) f_j^{(k)} f_h^{(k)}, & \text{for classification trees} \end{cases}$$



For an ordered response, let $s_1 < s_2 < \cdots < s_m$ be a system of numerical scores for categories. At a node t, consider the Generalized Gini impurity function:

$$i_{GG1}(t) = \sum_{h=1}^{m} \sum_{j=1}^{m} |s_h - s_j| f_h^{(t)} f_j^{(t)} \qquad (1)$$

$$i_{GG2}(t) = \sum_{h=1}^{m} \sum_{j=1}^{m} (s_h - s_k)^2 f_h^{(t)} f_j^{(t)} \qquad (2)$$

Prediction and error for nodes

Let $l_1, \ldots, l_T$ be the terminal nodes of a tree.

Node	l_1	l_2	$\cdots$	$\cdots$	l_T
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Prediction	$\hat{y}^{(l_1)}$	$\hat{y}^{(l_2)}$	$\cdots$	$\cdots$	$\hat{y}^{(l_T)}$
Error	$e^{(l_1)}$	$e^{(l_2)}$	$\cdots$	$\cdots$	$e^{(l_T)}$

- For observation i classified into node l, let $\hat{s}_i(l) = \hat{y}^{(l)}$ be the predicted score;
- The error $e^{(l)}$ entailed by the tree for observations classified into l can be either:
  - Total number of miss-classification:

$$e^{(l)} = R_{mr}(l) = \sum_{i \to l} \left(1 - I_{s_i}(\hat{s}_i(l)\right)$$

  - Total miss-classification cost:

$$e^{(l)} = R_{mc}(l) = \sum_{i \to l} |s_i - \hat{s}_i(l)|$$

- The total measure of predictive performance of a tree $\mathcal{T}$ with leaves $l_1 \ldots, l_T$ is given then by:

$$R_{mr}(\mathcal{T}) = \sum_i \left(1 - I_{s_i}(\hat{s}_i(\mathcal{T})\right) \text{ or } R_{mc}(\mathcal{T}) = \sum_i |s_i - \hat{s}_i(\mathcal{T})|$$

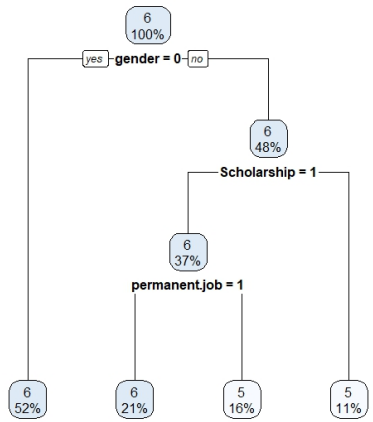
Setting hyper-parameters of the tree with pre-pruning

Pre-pruning *(stopping rules)*:

- Minimal sample size at node k necessary to look for candidate split (e.g. `minsplit` $= 50$; `minsplit` $= 0.3\,n, \ldots$);
- Minimal sample size of children nodes of a candidate split to accept it (e.g. textttminbucket $= 30$);
- Maximum depth of the tree (`maxdepth` $= 0$: only the root node; `maxdepth` $= 1$: (at most) the primary split; `maxdepth` $= 2$: (at most) one split down from the nodes of the primary split; $\ldots\ldots$;
- $\ldots\ldots$

```
library(rpart.plot)
```

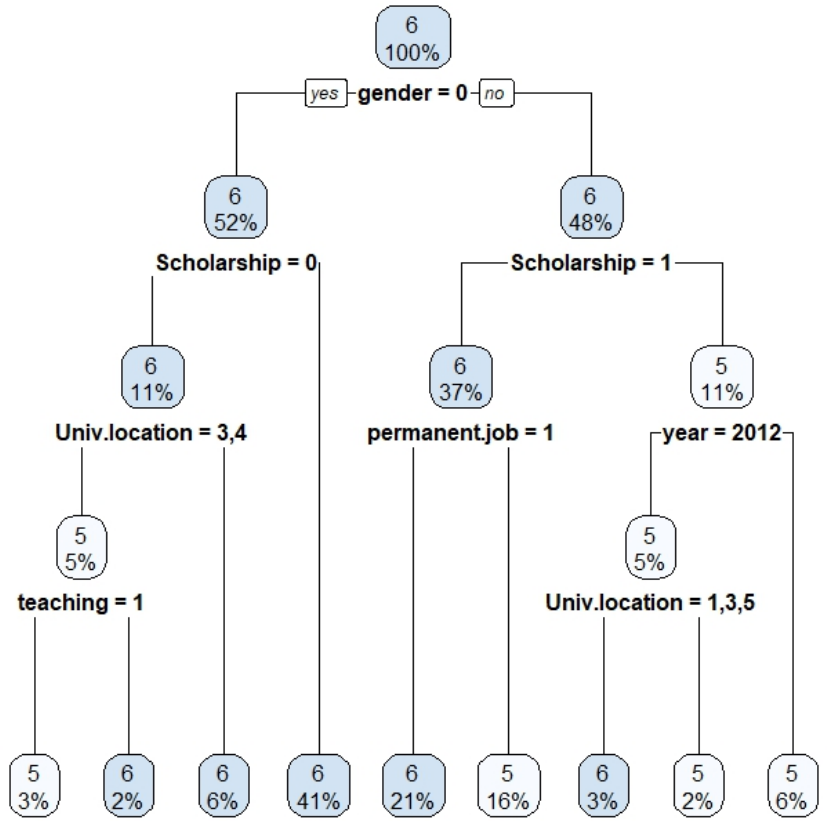
```
1) root 3830 4815 6
  2) gender=0 1992 2431 6 *
  3) gender=1 1838 2384 6
    6) Scholarship=1 1423 1782 6
     12) permanent.job=1 804  970 6 *
     13) permanent.job=0 619  791 5 *
    7) Scholarship=0 415  559 5 *
```



A deeper tree on Satisfaction for Professional Placement of PhDs

```
1) root 3830 4815 6
  2) gender=0 1992 2431 6
    4) Scholarship=0 425  586 6
      8) Univ.location=3,4 186  275 5
       16) teaching=1 115  183 5 *
       17) teaching=0 71   81 6 *
      9) Univ.location=1,2,5 239  303 6 *
    5) Scholarship=1 1567 1845 6 *
  3) gender=1 1838 2384 6
    6) Scholarship=1 1423 1782 6
     12) permanent.job=1 804  970 6 *
     13) permanent.job=0 619  791 5 *
    7) Scholarship=0 415  559 5
     14) year=2012 199  263 5
       28) Univ.location=1,3,5 117  151 6 *
       29) Univ.location=2,4 82  101 5 *
     15) year=2014 216  296 5 *
```

```
library(rpart.plot)
```



Selecting the best tree with post-pruning

(Cost Complexity) post-pruning:

- Instead of controlling the growth of the tree with pre-pruning rules, a very large tree $\mathcal{T}$ can be grown.
- Then, in order to correct for **overfitting**, some tree branches might be *pruned*;
- For each tree $\mathcal{T}$, consider the relative error improvement with respect to the root node ($\mathcal{T}^{(0)}$):

$$R^* = \frac{R(\mathcal{T})}{R(\mathcal{T}^{(0)})}.$$

- If $\mathcal{T}^{(-s)}$ denotes a given subtree obtained from $\mathcal{T}$ by pruning the splits s, the cost complexity parameter of pruning the branches (and nodes) not included in $\mathcal{T}^{(-s)}$ is defined as:

$$cp = \frac{R^*(\mathcal{T}^{(-s)}) - R^*(\mathcal{T})}{|\mathcal{T}| - |\mathcal{T}^{(-s)}|}$$

where the size $|\mathcal{T}|$ of a tree $\mathcal{T}$ is given by the number of its terminal nodes ($\texttt{nsplit}(\mathcal{T}) = |\mathcal{T}| - 1$);

- In general, if $s_1 \subset s_2$ are the sets of nodes that would be pruned from $\mathcal{T}$ to obtain subtrees $\mathcal{T}^{(-s_1)}$ and $\mathcal{T}^{(-s_2)}$, then the cost complexity of pruning nodes in $s_2 \setminus s_1$ would be:

$$c_p = \frac{R^*(\mathcal{T}^{(-s_2)}) - R^*(\mathcal{T}^{(-s_1)})}{|\mathcal{T}^{(-s_1)}| - |\mathcal{T}^{(-s_2)}|}$$

- Remark that pruned trees entail larger errors R^*, so $cp > 0$

Ctree: conditional inference trees

- Trees built with CART do not imply any conclusion on statistically significant differences of splitting variables at children nodes. This could cause overfitting and biased trees (in which variables with many possible split-points are more likely to occur in the partitioning process). Variable selection and split-point selection are parallel procedures.

Conditional inference Trees are *unbiased* trees so that:

- First, a global permutation test of independence between response and predictors is run. If significant, then permutation tests are run individually for each predictor: the one with most significant association with response is retained *(variable selection phase)*.
- For the selected predictor, permutation tests are run to identify the best split-point.
- The procedure works for any kind of response: for ordinal variable, it requires that numeric scores are assigned to categories.

Hothorn, T., Hornik, K., Zeileis, A. (2019). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, Taylor & Francis, 2006, 15, 651-674

Hothorn, T., Zeileis, A. (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research*, 16:3905-3909.

Conditional inference trees: library(partykit)

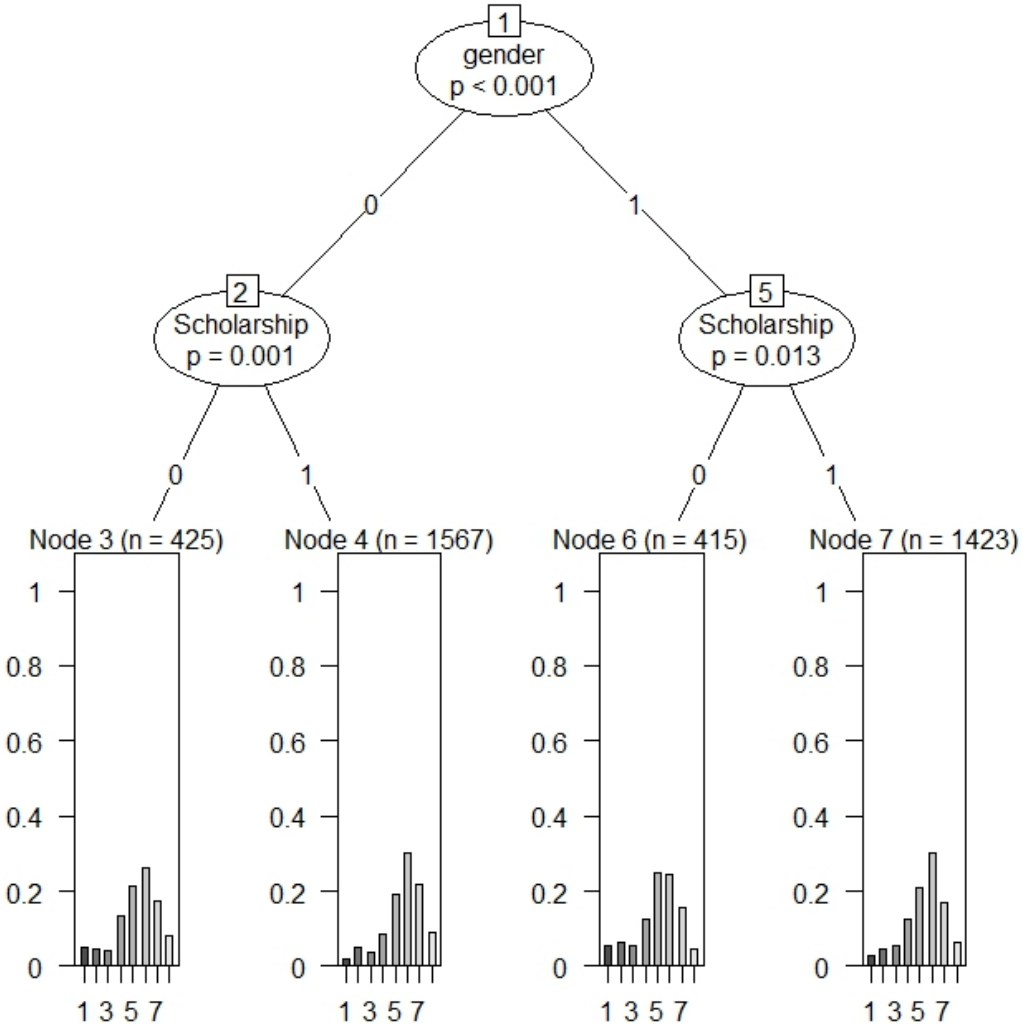


Figure: CTREE for satisfaction for PhD placement

A proposal for Quantile ANOVA tree

- Quantiles are the location measures most suitable to summarize rating distributions, as they are structured on the order among categories and their computation does not require any numerical scoring.
- IDEA: implement a Quantile-ANOVA tree for ratings where - at each step - significant differences in possibly many quantiles drive the partitioning process;
- Quantile Anova could be also used to perform *multi-group* analysis on the terminal nodes.


Wilcox, R. R., Erceg-Hurn, D., Clark, F., & Carlson, M. (2014). Comparing two independent groups via the lower and upper quantiles. *Journal of Statistical Computation and Simulation*, **84**, 1543–1551.

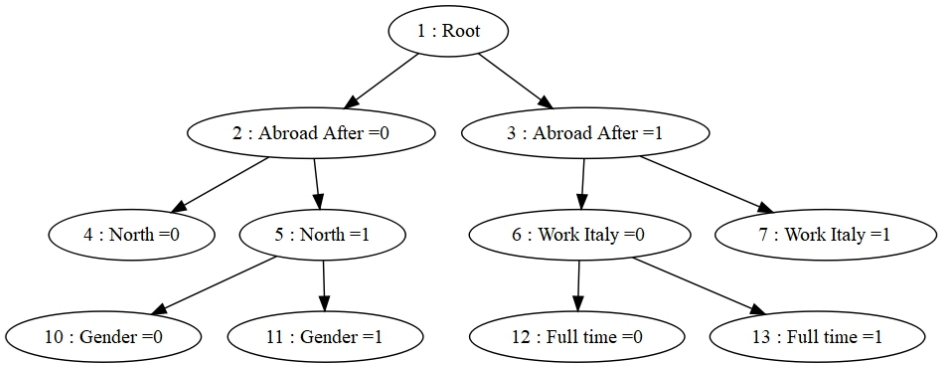

Wilcox, R.R.(2017). Introduction to Robust Estimation & Hypothesis


Mair, P., Wilcox, R.R. (2019). Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*, https://doi.org/10.3758/s13428-019-01246-w.


Simone, R., Davino, C., Vistocco, D., Tutz, G. (2023). Quantile-based decision trees for ordinal rating responses. (preprint)

Drivers of dissatisfaction of Italian PhDs

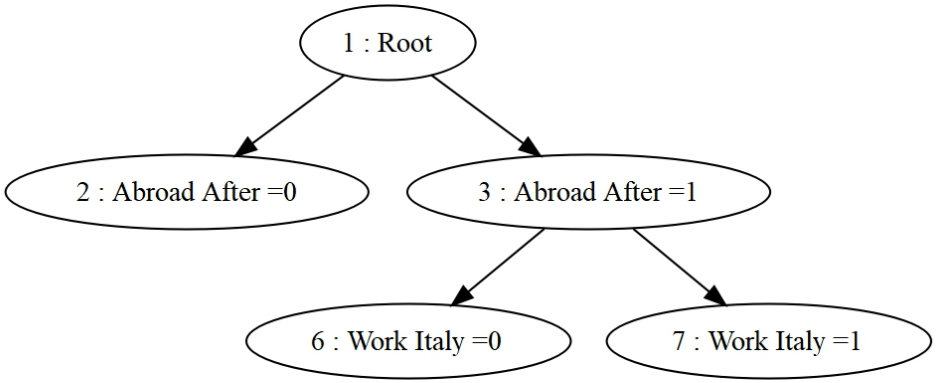
Lower quantile tree: $S(\boldsymbol{q}) = \{q_{0.1}, q_{0.25}\}$



	Root	Terminal Nodes					
		4	7	10	11	12	13
$q_{0.1}$	3	2	2	3	2	5	5
$q_{0.25}$	5	4	4	5	4	5	6

Drivers of satisfaction of Italian PhDs

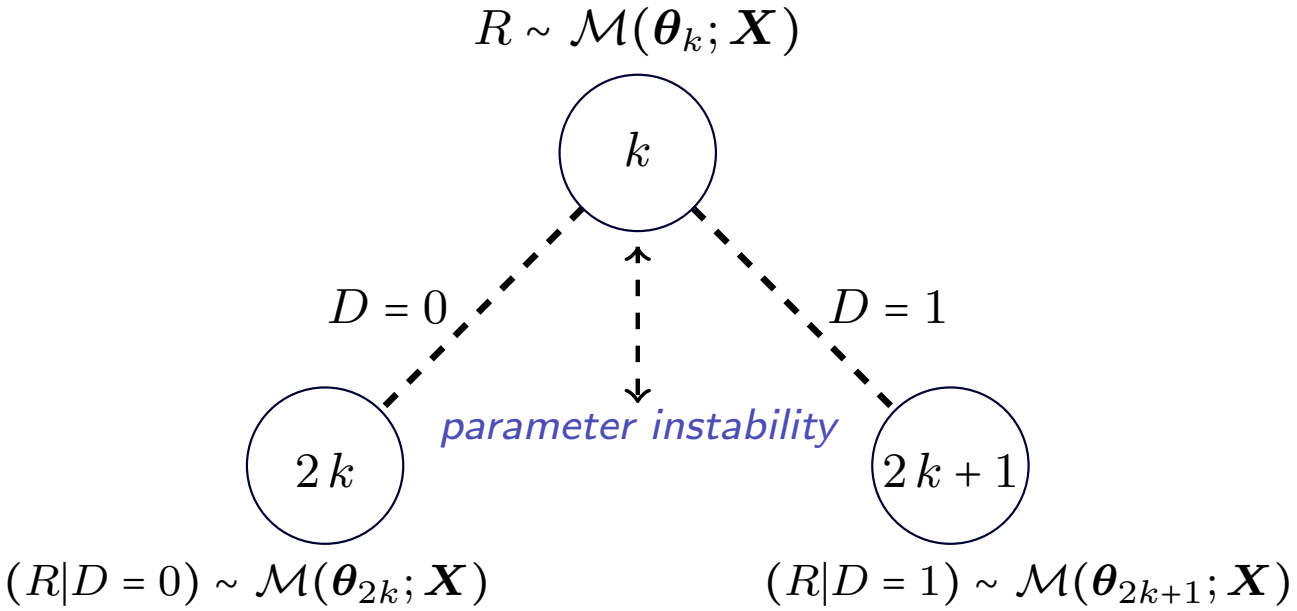
Upper quantile tree: $S(\boldsymbol{q}) = \{q_{0.75}, q_{0.9}\}$



	Root	Terminal Nodes		
		2	6	7
$q_{0.75}$	7	6	7	6
$q_{0.9}$	7	7	8	7

Model-based Trees

- Given a model $\mathcal{M}(\boldsymbol{\theta}; \boldsymbol{X})$ for the response R, a binary tree is grown to derive respondents' profiles by means of recursive partitioning on the basis of splitting variables;
- at each step, M-fluctuation tests are used to identify the candidate splitting variable that entails the higher instability to parameters $\boldsymbol{\theta}_k$ when entering the model;
- Then, the split point that maximizes an objective function is selected.



Zeileis, A., Hothorn, T., Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, **17**, 492 – 514.

Hothorn, T., Zeileis, A. (2015). `partykit`: A Modular Toolkit for Recursive Partytioning in R. Journal of Machine Learning Research, 16, 3905-3909. URL http://jmlr.org/papers/v16/hothorn15a.html

Latent variable approach: Cumulative link models

- Let $R_i^\star$ denote the underlying (continuous) latent variable for the i-th subject, and let R_i be the ordinal score marked by the i-th respondent to an item of a questionnaire for $i = 1, \dots, n$. If $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_m = +\infty$, then:

$$\alpha_{r-1} < R_i^\star \leq \alpha_r \iff R_i = r, \qquad r = 1, \dots, m$$

(Benchmark bibliography)


McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109–142.


Agresti, A. (2010), *Analysis of Ordinal Categorical Data*, Wiley Series in Probability and Statistics.


Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

- The paradigm assumes a standard regression model on the latent trait:

$$R_i^\star = \mathbf{t}_i \, \boldsymbol{\beta} + \epsilon_i,$$

with $p \geq 1$ subjects' covariates $\mathbf{t}_i$ and regression coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$

- A link is set between the cumulative distribution and the linear predictor:

$$Pr(R_i \leq r | \boldsymbol{\theta}, \mathbf{t}_i) = Pr(R_i^\star \leq \alpha_r | \boldsymbol{\theta}, \mathbf{t}_i) = Pr(\epsilon_i \leq \alpha_r - \mathbf{t}_i \, \boldsymbol{\beta}) = F_\epsilon\big(\alpha_r - \mathbf{t}_i \, \boldsymbol{\beta}\big)$$

MOB on stress

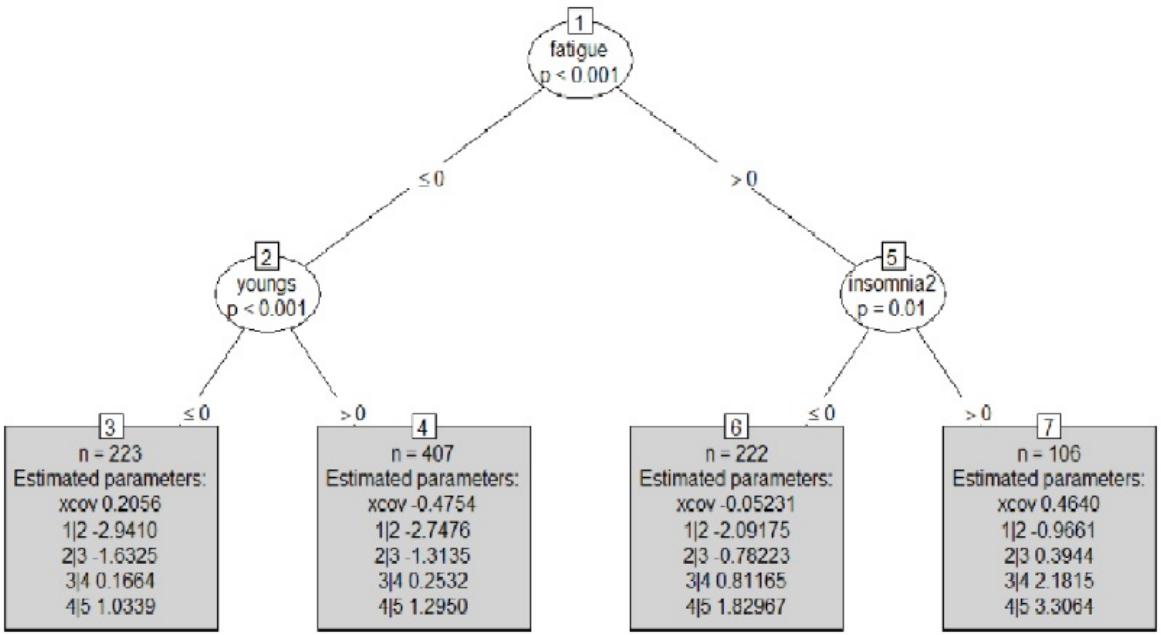


FIGURE: MOB tree with ordinal logits ($\mathcal{M}$: Stress ~ Gender)

Models on the discrete support

A different approach foresees to model ordinal response variables for preference data directly on the discrete support ($\{c_1 \prec c_2 \prec \cdots \prec c_m\}$) rather than on the continuous latent scale

Jenkins S.P. (2020). Comparing distributions of ordinal data. *The Stata Journal*, 20(3), 505–531.

In this case, for the observed sample $(r_1, \ldots, r_n)$ with relative frequency distribution $(f_1, \ldots, f_m)$, the fitting result is directly a probability model $(p_1(\boldsymbol{\theta}), \ldots, p_m(\boldsymbol{\theta}))$, where possibly $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_i$ depending on subjects' covariates

- Discretized Beta distribution

Taverne, C. and Lambert, P. (2014). Inflated Discrete Beta Regression Models for Likert and Discrete Rating Scale Outcomes, *arXiv:1405.4637v1*, 19 May, 2014.

Ursino M (2014) *Ordinal Data: a New Model with Applications*. Ph.D. Thesis, XXVI cycle, Polytechnic University of Turin, Turin

Ursino, M. and Gasparini, M. (2018). A new parsimonious model for ordinal longitudinal data with application to subjective evaluations of a gastrointestinal disease, *Statistical Methods in Medical Research*, **27**(5), 1376–1393.

Simone, R. (2022) On finite mixtures of Discretized Beta model for ordered responses. *TEST* 31, 828–855.

Models on the discrete support

A different approach foresees to model ordinal response variables for preference data directly on the discrete support ($\{c_1 \prec c_2 \prec \cdots \prec c_m\}$) rather than on the continuous latent scale

Jenkins S.P. (2020). Comparing distributions of ordinal data. *The Stata Journal*, 20(3), 505–531.

In this case, for the observed sample $(r_1, \ldots, r_n)$ with relative frequency distribution $(f_1, \ldots, f_m)$, the fitting result is directly a probability model $(p_1(\boldsymbol{\theta}), \ldots, p_m(\boldsymbol{\theta}))$, where possibly $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_i$ depending on subjects' covariates

- Inverse Hypergeometric distribution and its mixture

D'Elia A. (2003). Modelling ranks using the inverse hypergeometric distribution, *Statistical Modelling*, 3, 65–78

Simone R., Iannario M. (2018). Analysing sport data with clusters of opposite preferences. *Statistical Modelling*, **18**(5-6), 1–20

Models on the discrete support

A different approach foresees to model ordinal response variables for preference data directly on the discrete support ($\{c_1 \prec c_2 \prec \cdots \prec c_m\}$) rather than on the continuous latent scale

Jenkins S.P. (2020). Comparing distributions of ordinal data. *The Stata Journal*, 20(3), 505–531.

In this case, for the observed sample $(r_1, \ldots, r_n)$ with relative frequency distribution $(f_1, \ldots, f_m)$, the fitting result is directly a probability model $(p_1(\boldsymbol{\theta}), \ldots, p_m(\boldsymbol{\theta}))$, where possibly $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_i$ depending on subjects' covariates

- Binomial distribution

Allik J (2014) A mixed-binomial model for Likert-type personality measure. *Frontiers in Psychology* 5:1–13

Pinto da Costa JF, Alonso H, Cardoso JS (2008) The unimodal model for the classification of ordinal data. *Neural Networks*, **21**, 78–91. *Corrigendum* in: (2014). *Neural Networks*, **59**, 73–75

Zhou H, Lange K (2009) Rating movies and rating the raters who rate them. *The Amer Stat*, 63:297–307

Mixture models with uncertainty for ordinal variables

The class of CUB mixture models [2] for ordinal variables $(R_1, \ldots, R_n)$ is grounded on the specification of an *uncertainty* and a feeling component:

$$Pr\left(R_i = r \mid \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\theta}\right) = \pi_i \, b_r(\xi_i \mid \boldsymbol{w}_i) + \left(1 - \pi_i\right) \frac{1}{m}, \qquad r = 1, \ldots, m, \; i = 1, \ldots, n$$

Shifted Binomial:

$$b_r(\xi_i) = \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1}$$

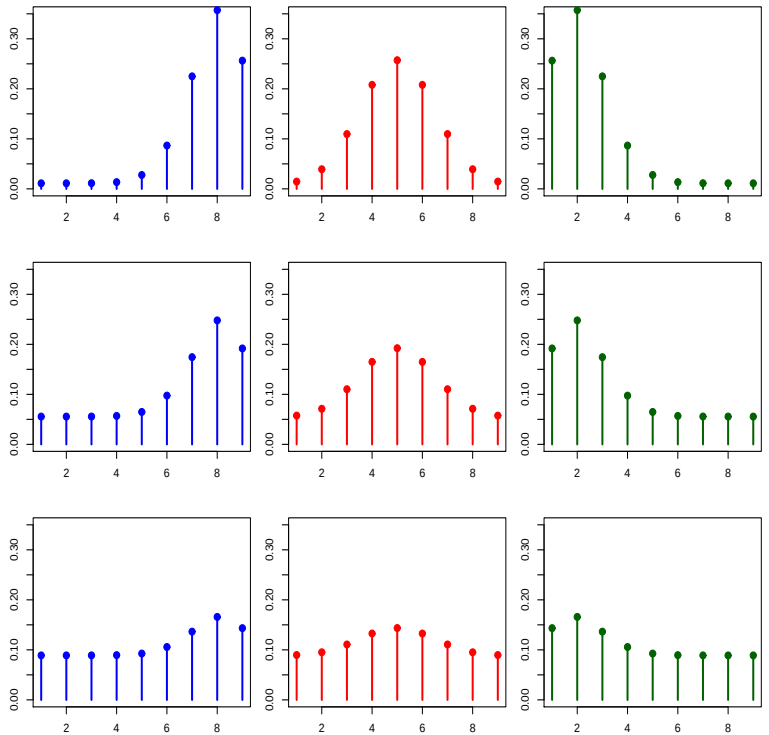
Systematic components:

$$\text{logit}\left(\pi_i\right) = \boldsymbol{x}_i \boldsymbol{\beta}$$

$$\text{logit}\left(\xi_i\right) = \boldsymbol{w}_i \boldsymbol{\gamma}$$

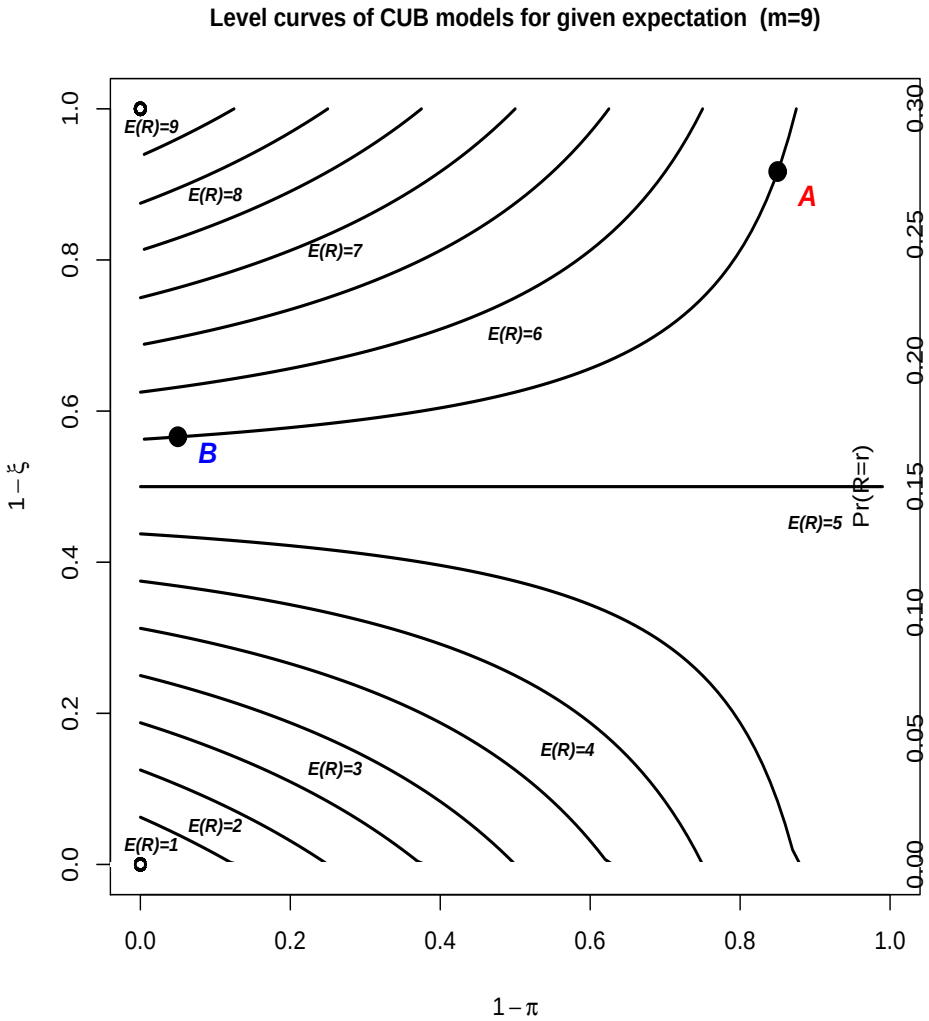
No covariate:

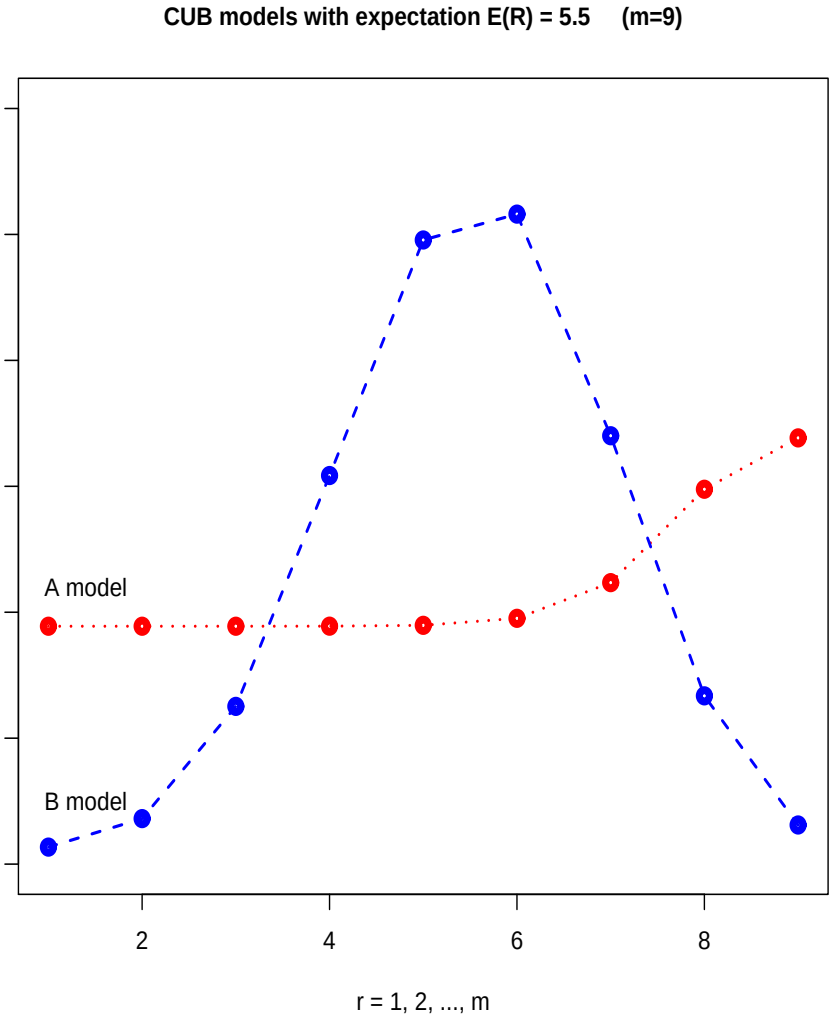
$$\pi_i = \pi \in (0, 1], \qquad \xi_i = \xi \in [0, 1]$$



[2] D'Elia & Piccolo (2005). Piccolo and Simone (2019)

CUB MODELS VISUALIZATION: CONTOUR PLOT OF EXPECTATION





Role and Meaning of Uncertainty [a]

[a] Piccolo and Simone 2019

Sampling from CUB (π, ξ_i), for $\pi \in (0, 1]$:

$$\text{logit}(\xi_i) = \gamma_0 + \gamma_1 D_i \qquad \text{logit}(Pr(R_i \leq r | D_i)) = \alpha_r - \delta_1 D_i$$

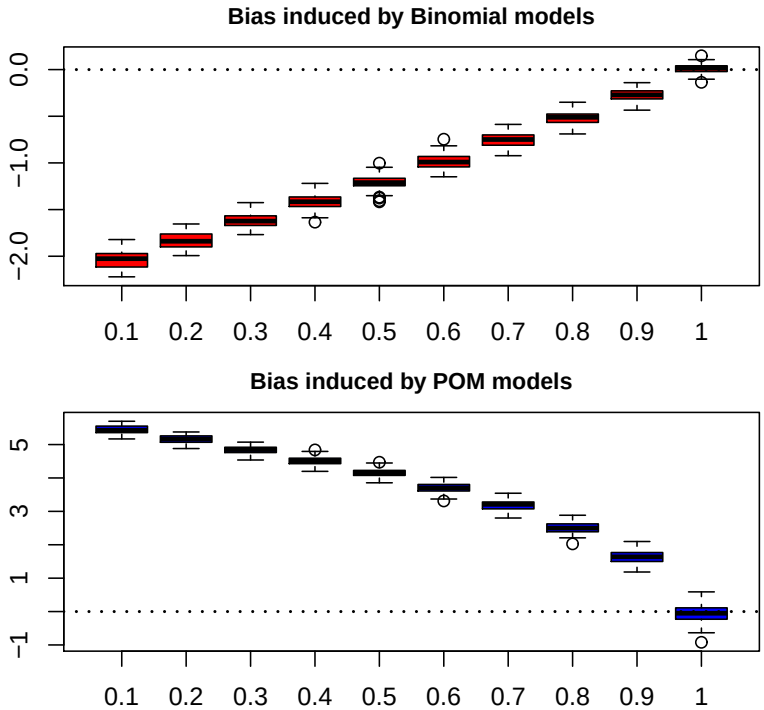
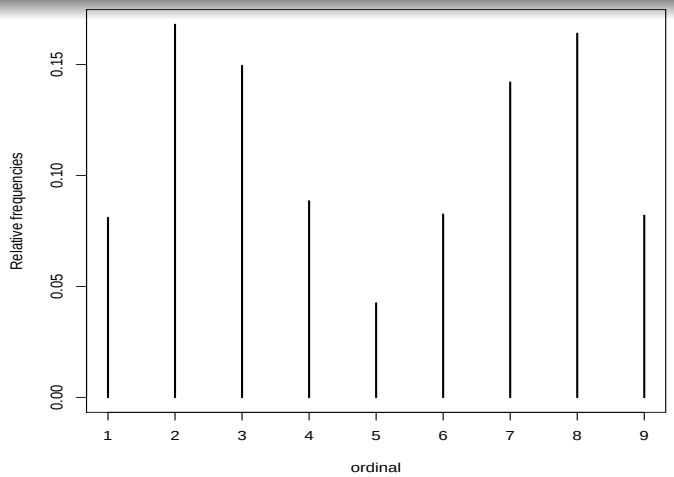


Figure: Bias in the estimation of γ_1 for Binomial model (top) and in the estimation of δ_1 in POM (bottom) for data generated from CUB with varying values of π (abscissa)

CUB MODELS AND VARIABLE SELECTION



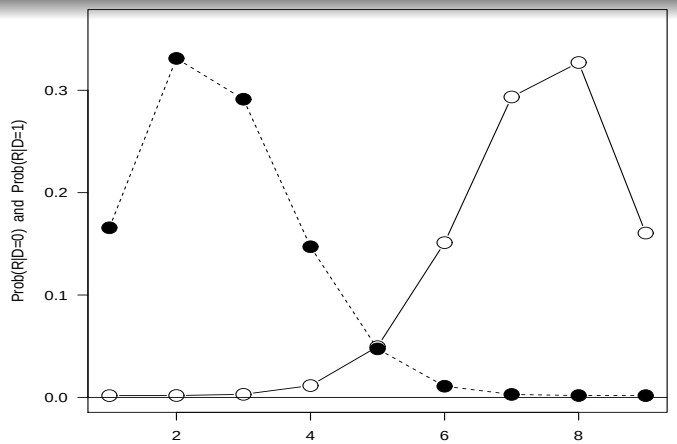


Figure shows the simulated and estimated distributions (conditional to $D_i = 0, 1$, respectively) of the shifted Binomial model ($m = 9$):

$$\begin{cases} Pr(R_i = j) = \binom{8}{j-1} \xi_i^{8-j} (1-\xi_i)^{j-1}; \\ logit(\xi_i) = -1.362 + 2.744\, D_i; \end{cases} \qquad j = 1, 2, \ldots, 9; \;\; i = 1, 2, \ldots, n.$$

- Combined Backward - Forward selection for feeling and uncertainty parameters;
- Best-subset via accelerated EM algorithm:

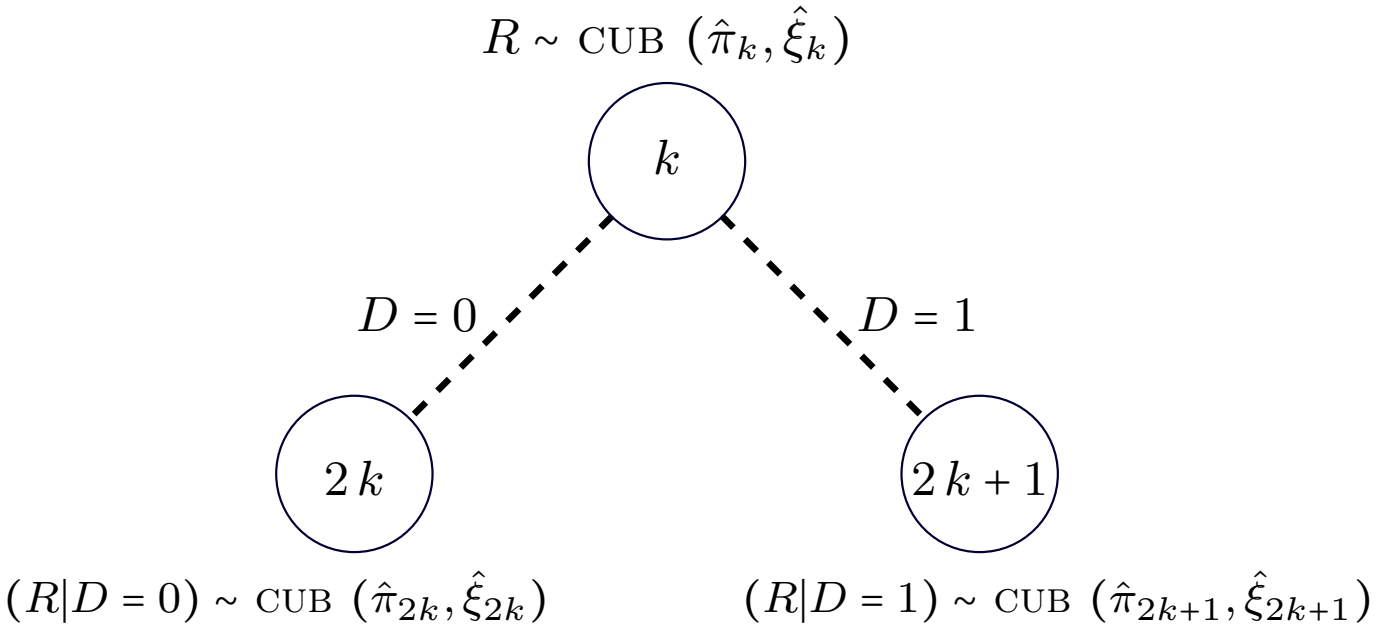
  Simone R. (2021). An accelerated EM algorithm for mixture models with uncertainty for rating data. Computational Statistics, **36**:691–714

- Group Lasso-regularization (Schneider, Pößnecke, Tutz: `https://epub.ub.uni-muenchen.de/68452/`)

CUBREMOT (CUB Regression Model Trees)

- An explanatory dummy among the available ones is identified which *best* splits the data into two groups $R|D=0$ and $R|D=1$ (according to a *splitting criterion*)
- At node k with n_k observations, let $R \sim \text{CUB}\ (\pi_k, \xi_k)$ $(m > 3)$;
- If D is a significant dummy D to explain *uncertainty* and/or *feeling*, then:

$$\text{logit}(\pi_k) = \beta_0^{(k)} + \beta_1^{(k)} D, \qquad \text{logit}(\xi_k) = \gamma_0^{(k)} + \gamma_1^{(k)} D$$



Cappelli, C., Simone, R., Di Iorio, F. (2019). CUBREMOT : a model-based tree for ordinal responses. *Expert Systems with Applications*, **124**:39–49.

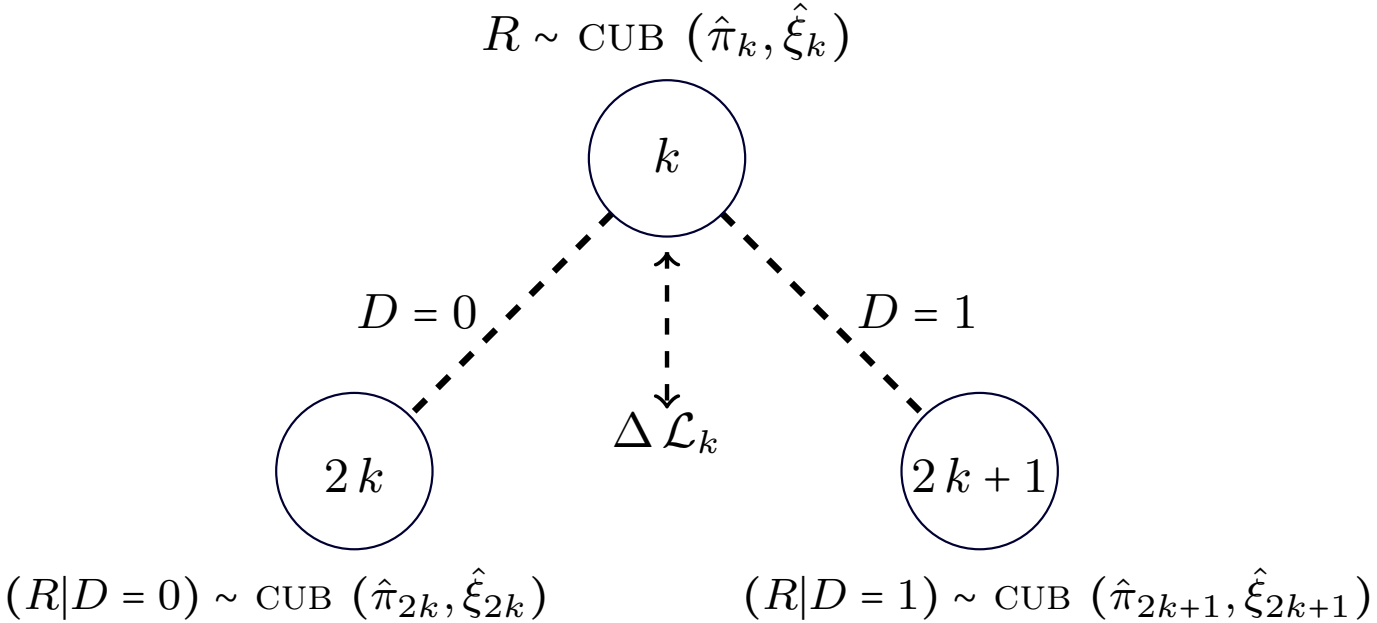
Splitting criteria

Deviance-based decision rule

At any node k, choose the split induced by the (significant) covariate D that maximizes:

$$\Delta \mathcal{L}_k = \left(\mathcal{L}_{n_0}(\hat{\pi}_{2k}, \hat{\xi}_{2k}) + \mathcal{L}_{n_1}(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})\right) - \mathcal{L}_n(\hat{\pi}_k, \hat{\xi}_k)$$

where $\mathcal{L}_n(\hat{\pi}_k, \hat{\xi}_k)$ is the log-likelihood of a CUB fit on node k with n observations.



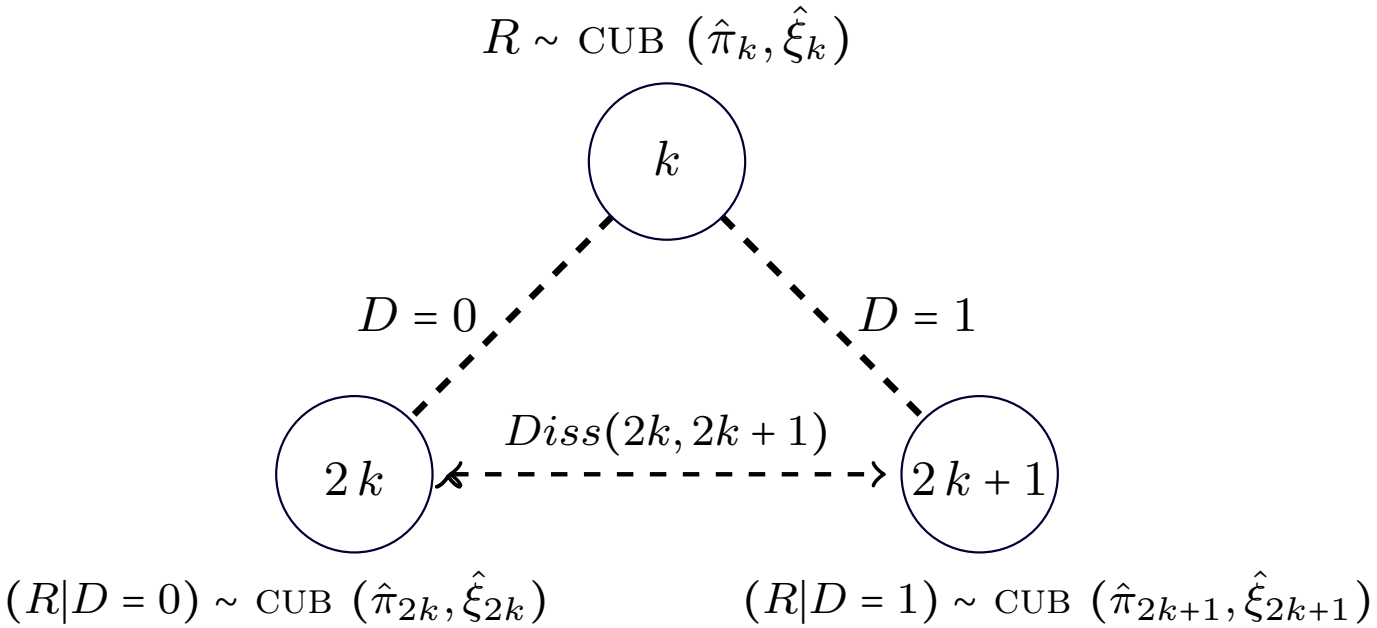
Splitting criteria

Dissimilarity-based decision rule

At any node k, choose the split that implies the maximum *dissimilarity* between child nodes:

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

where $\hat{p}^{(2k)}$ and $\hat{p}^{(2k+1)}$ are the estimated probability distributions for the child nodes.



Choosing the best tree

If $l_1, \ldots, l_T$ are the terminal nodes (leaves) of a CUBREMOT grown according to a given criterion, with sizes $n(l_1), \ldots, n(l_T)$, and if

$$Diss(l_j) = \frac{1}{2} \sum_{r=1}^{m} |\hat{p}_r^{(l_j)} - f_r^{(l_j)}|, \quad j = 1, \ldots, T$$

with $(f_1^{(l_j)}, \ldots, f_m^{(l_j)})$ and $(p_1^{(l_j)}, \ldots, p_m^{(l_j)})$ observed frequency and estimated probability distributions at the j-th leaf, resp., then choose the tree with minimum *weighted dissimilarity* at terminal nodes:

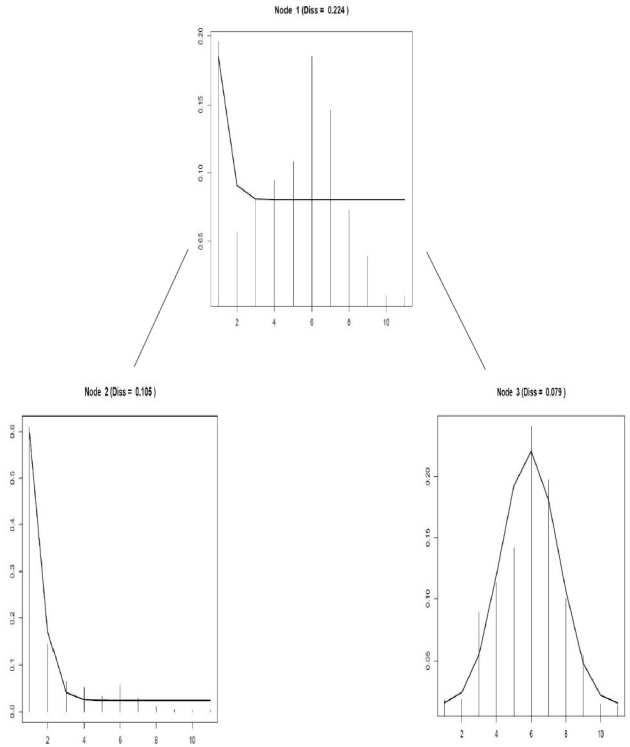
$$Diss_w = \sum_{j=1}^{T} \frac{n(l_j)}{n} Diss(l_j) \; = \; min! \tag{3}$$

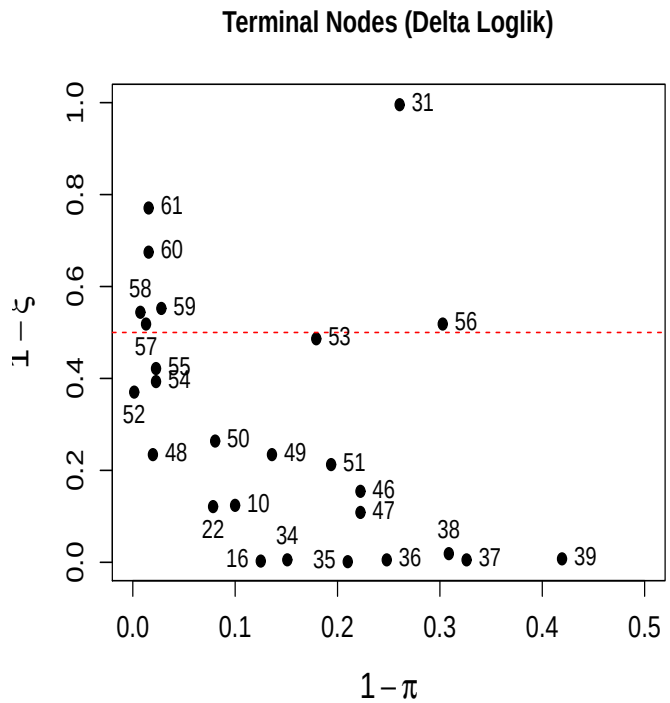
where $n = \sum_{j=1}^{T} n(l_j)$ is the total number of observations.

Visualization for CUBREMOT

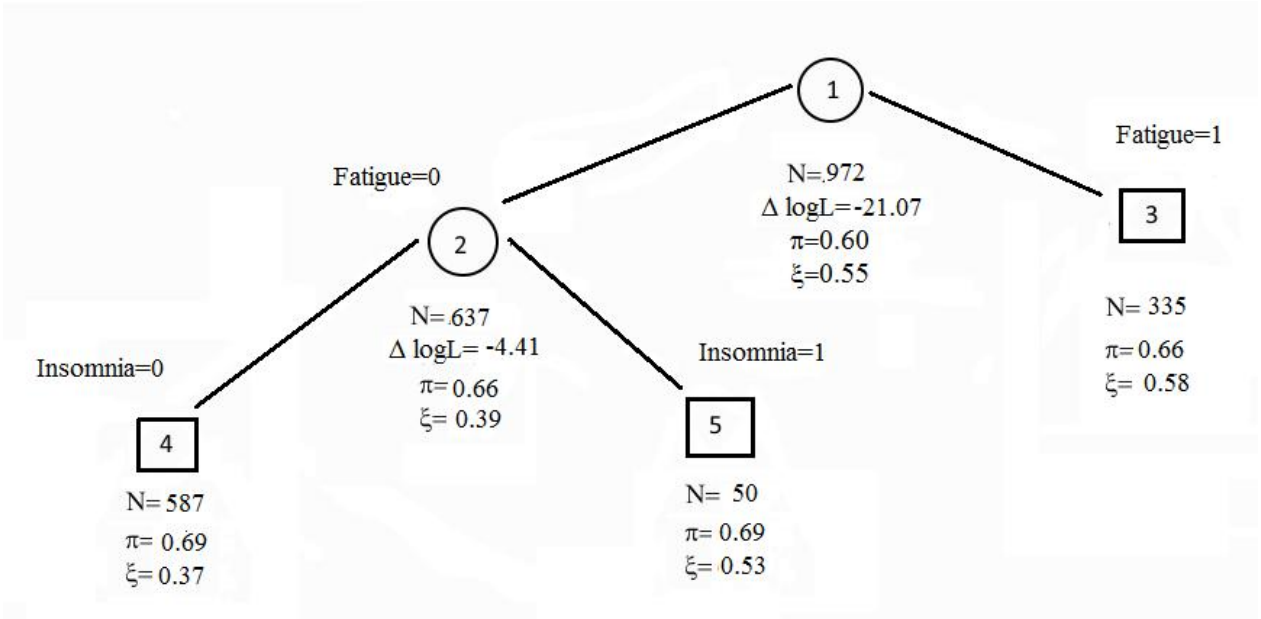
Effective graphical displays are available for CUBREMOT :

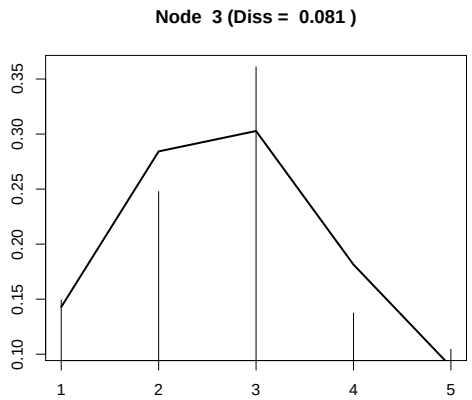
- Plot of observed and fitted distributions of the response in each internal and/or terminal node
- Scatter plot of the estimated CUB models at terminal nodes as points in the parameter space

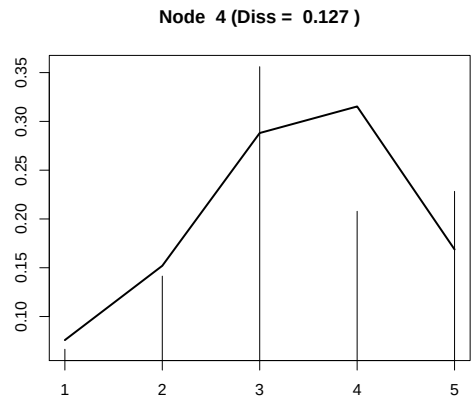


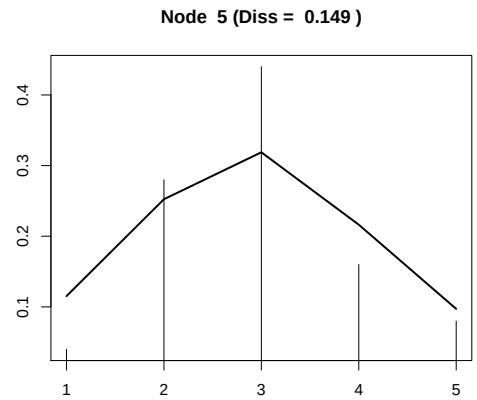


The stress tree: Log-likelihood splitting rule









Shelter effect

If c denotes the *shelter* category, let

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

$R \sim \text{CUB}_{she}(\pi^\star, \xi, \delta)$, with shelter at c, if:

$$Pr(R = r|\boldsymbol{\theta}^\star) = (1-\delta)\left(\pi^\star b_r(\xi) + (1-\pi^\star)\frac{1}{m}\right) + \delta\, D_r^{(c)}$$

Possibly, with subjects covariates $\mathbf{v}_i$:

$$\text{logit}(\delta_i) = \boldsymbol{v}_i\, \boldsymbol{\omega}$$

Corduas M., Iannario M., Piccolo D. (2009). A class of statistical models for evaluating services and performances, in: M.Bini *et al.* (eds.): *Statistical methods for the evaluation of educational services and quality of products*. Contribution to Statistics, Physica-Verlag, Springer, Berlin Heidelberg, pp.99–117

Iannario M. (2012). Modelling *shelter* choices in a class of mixture models for ordinal responses. *Stat Meth Appl*, **21**:1–22.

CUBREMOT -S...2

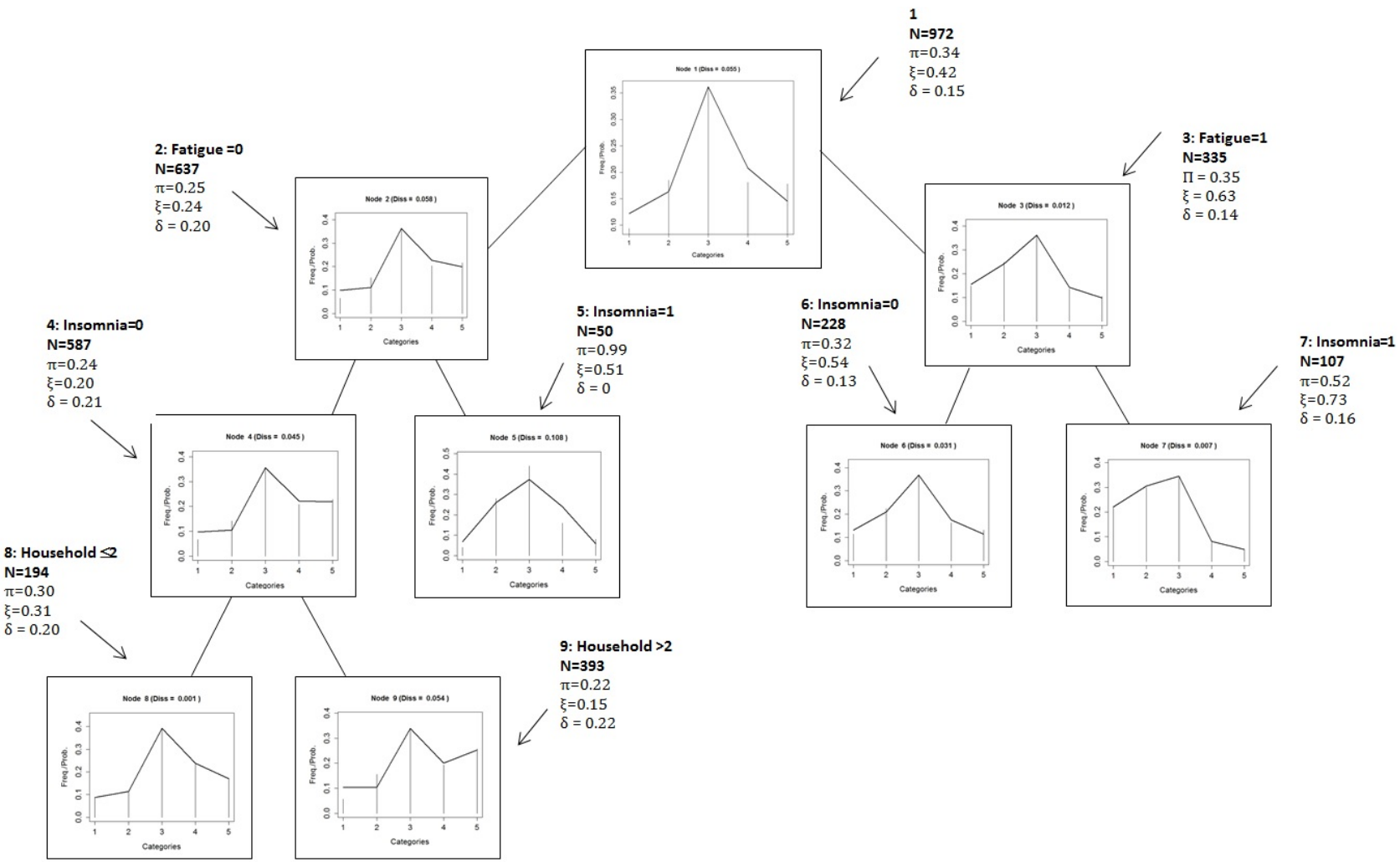


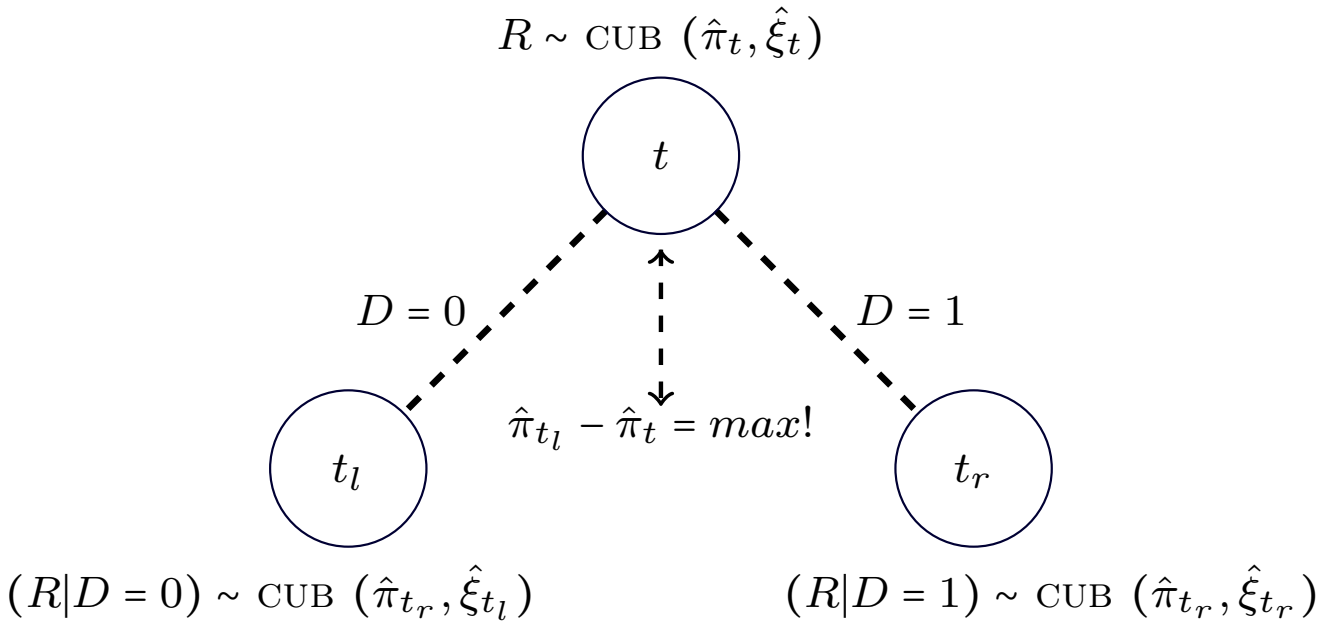
FIGURE: CUBREMOT -S grown with the log-likelihood splitting criterion

The uncertainty tree

- Consider CUB *uncertainty* as *impurity* in growing a tree.
- CUB models with dummy covariates only for feeling are tested on node t to obtain candidate splits so that:

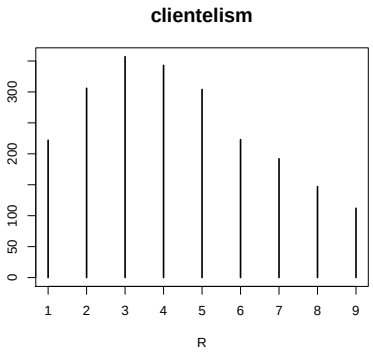
$$logit(\hat{\xi}_{t_l}) = \hat{\gamma}_0^{(t)}, \qquad logit(\hat{\xi}_{t_r}) = \hat{\gamma}_0^{(t)} + \hat{\gamma}_1^{(t)},$$

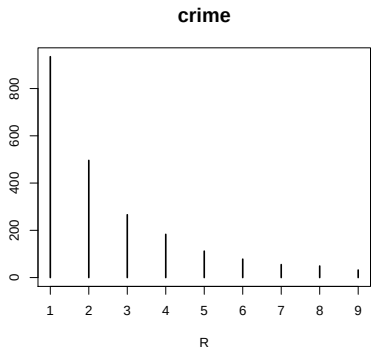
and constant uncertainty parameters $\hat{\pi}_{t_l} = \hat{\pi}_{t_r}$.

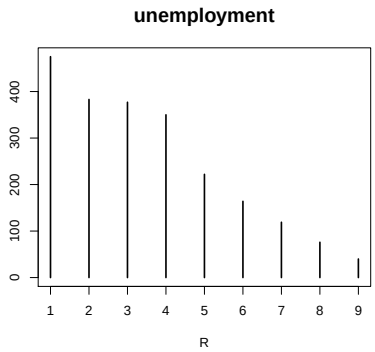


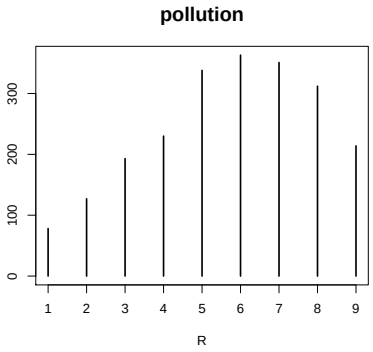
Simone, R., Cappelli, C., Di Iorio, F. (2019). Modelling marginal ranking distributions: the uncertainty tree, *Pattern Recognition Letters*, **125**: 278–288.

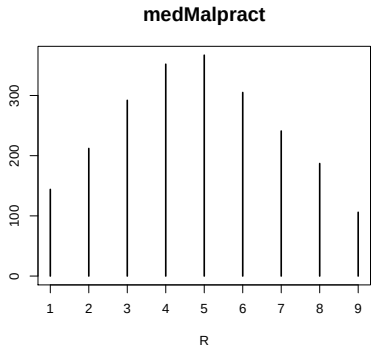
Data example 3: Marginal rankings for metropolitan emergencies in Naples (2007)

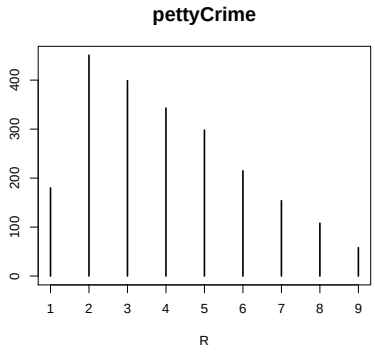


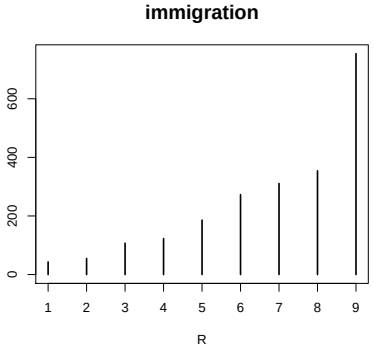


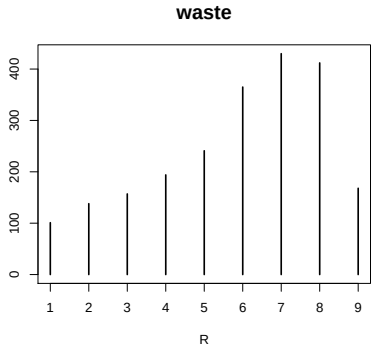


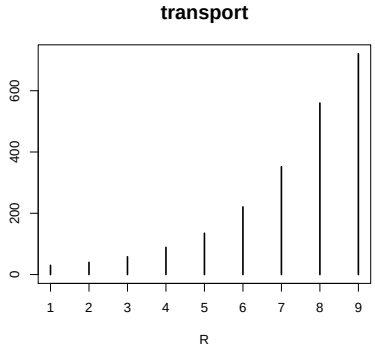




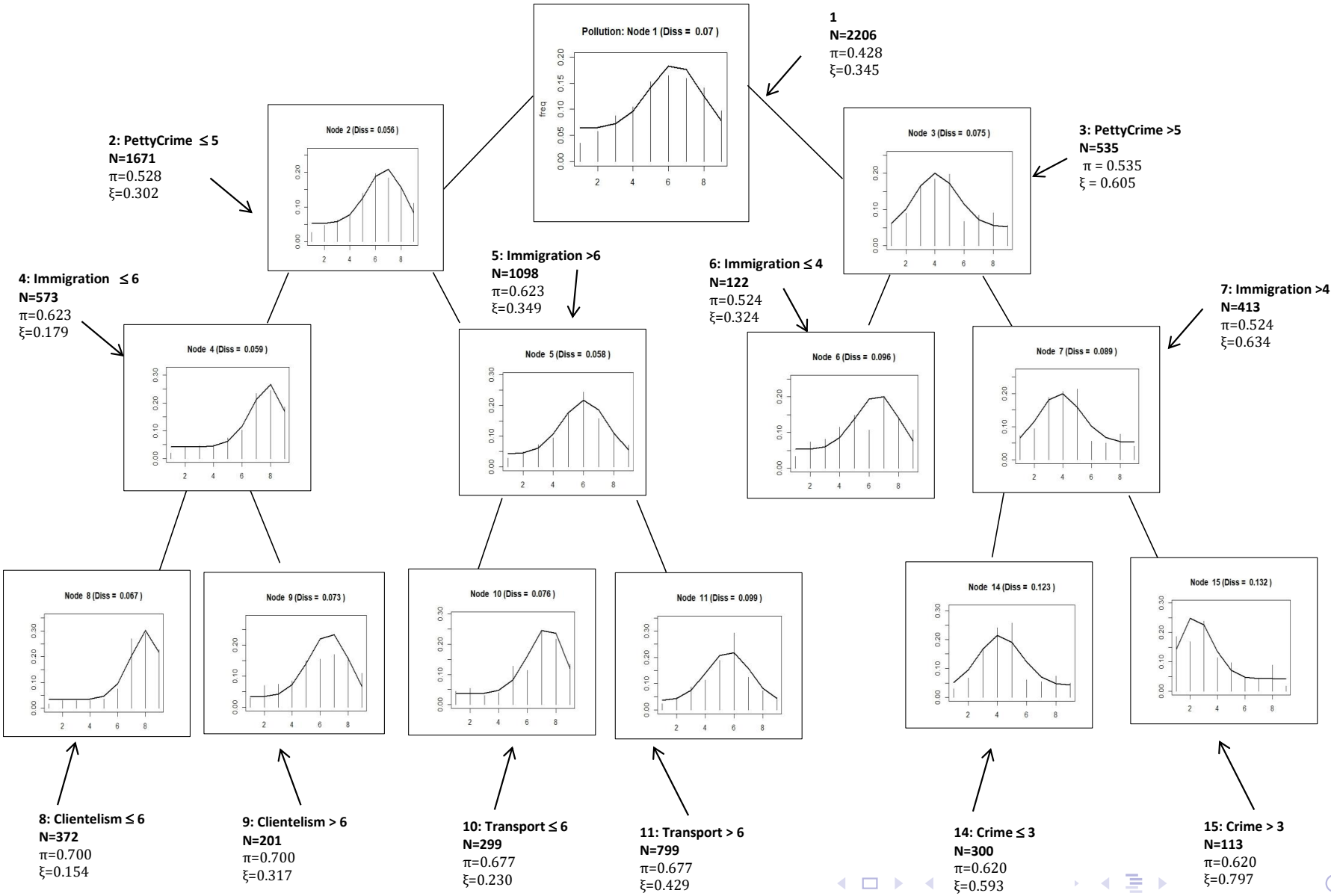








Pollution: uncertainty CUBREMOT



CUBREMOT versus MOB

- CUBREMOT allows to customize the splitting criterion in order to derive meaningful response profiles, parameterized in terms of featuring parameters;
- In principle, it does not require covariate specification for the maintained model, yet the procedure can be easily extended so that the baseline model at each node is a CUB model with covariates;
- Different visualization tools are available;
- The forthcoming `R` package will tackle possible bias issues in the selection of the splitting variable by allowing separate variable selection step and split-point choice;
- Possibility to implement flexible trees.....

Residual Diagnostics for ordinal models

Liu, D. and Zhang, H. (2018). Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach. *Journal of the American Statistical Association*, **113**, 845 – 854.

- For models of the form $Y \sim F_a(y; X, \theta)$, where $F_a(\cdot)$ is the cumulative distribution function of the *assumed* model, the Authors advocate a *jittering* (on the probability scale) approach on the probability scale and define a surrogate variable S by conditionally sampling from a Uniform distribution:

$$S|Y = y \sim \mathcal{U}(F_a(y-1), F_a(y)), \quad y = 1, \ldots, m$$

- Then, the residuals for the model fit can be defined as:

$$R = S - \mathbb{E}[S|\eta]$$

under the null that the assumed model is correctly specified, where η denotes the available information set.

- Liu and Zhang (2018) proved that $R|X \sim \mathcal{U}(-\frac{1}{2}, \frac{1}{2})$ if the assumed model is correctly specified.
- Test for Uniformity: general (Kolmogorov-Smirnov, Cramer - von Mises, ...), specific (Greenwood, Quesenberry-Miller (1977), ...)

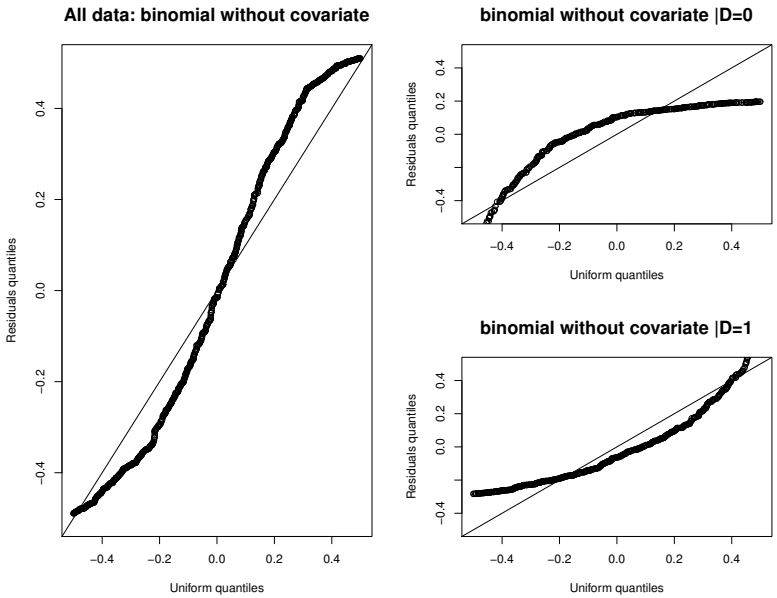
Quesenberry, C.P. and Miller, F.L.Jr. (1977). Power studies of some tests for uniformity. *Journal of the Statistical Computation and Simulation*, **5**(3), 169 – 191.

Neglecting mixed population

Generate data according to $R \sim Bin(\xi_i), \text{logit}(\xi_i) = \gamma_0 + \gamma_1 D_i$, with $m = 7$

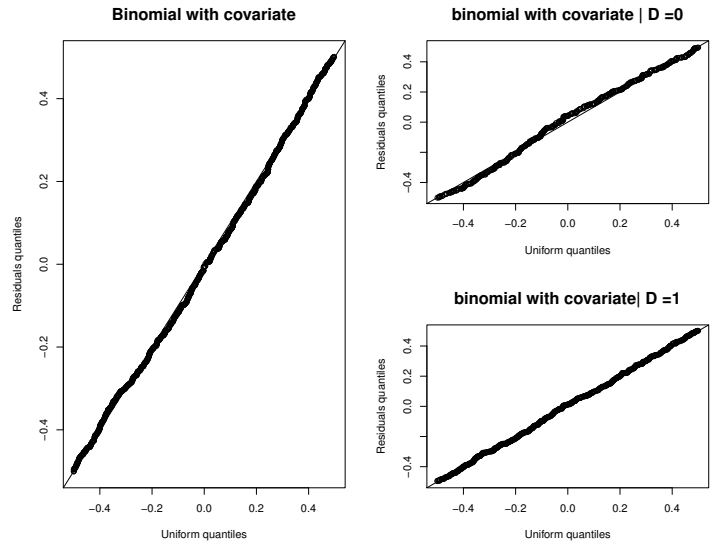
Assumed model:

$$R \sim Bin(\xi)$$



Assumed model:

$$R \sim Bin(\xi_i),\ \text{logit}(\xi_i) = \gamma_0 + \gamma_1 D_i$$



Simone, R. (2023). Uncertainty Diagnostics of Binomial Regression Trees for Ordered Rating Data. *Journal of Classification*, doi:10.1007/s00357-022-09429-5

Further applications on tree diagnostics

- Select the most performing splitting criterion;
- Set optimal values for pre-pruning conditions (also for general MOB trees)
- Implement flexible uncertainty trees:

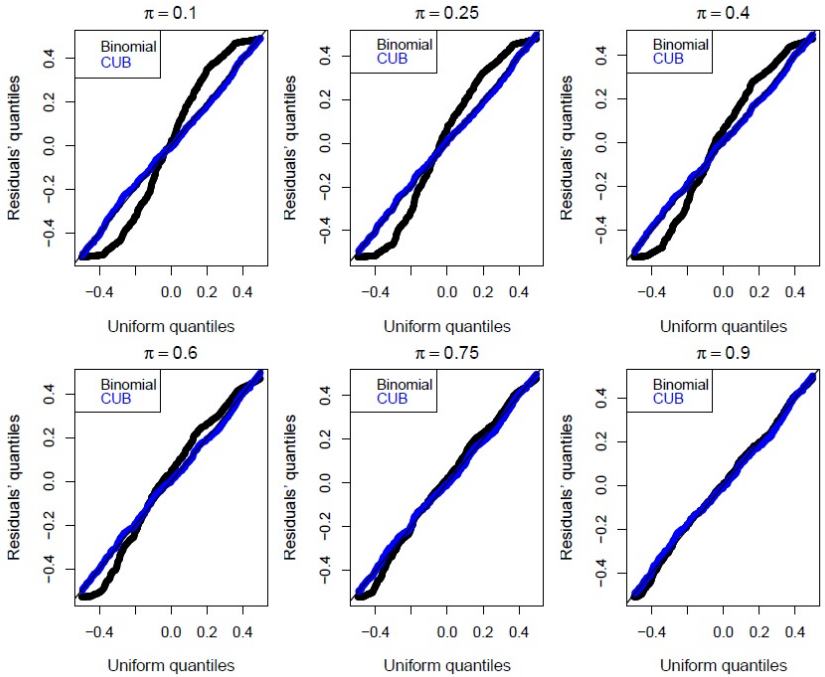


Figure: QQ-plot of surrogate residuals for Binomial and CUB fit to data generated according to CUB $(\pi, \xi = 0.7)$, with $m = 7$, for varying $\pi \in (0, 1]$

Simone, R. (2023). Uncertainty Diagnostics of Binomial Regression Trees for Ordered Rating Data. *Journal of Classification*, doi:10.1007/s00357-022-09429-5

Flexible uncertainty tree on satisfaction for PhD experience

Table: Summarizing results for the local model selection on the (deviance) CUBREMOT for Ph.D. overall satisfaction

Node	Best Model	$\hat{\pi}$	$1-\hat{\xi}$	$\hat{\delta}$	$\hat{\phi}$	shelter	Split with	Diss(CUB , $\mathbf{f}$)	Diss(Best, $\mathbf{f}$)
1	CUB +she	0.52	0.664	0.037	-	6	L: research= 0, R: research= 1	0.038	0.027
2	CUB +she	0.39	0.604	0.040	-	6	L: stem= 0; R: stem= 1	0.039	0.025
3	CUB +she	0.67	0.683	0.016	-	2	L: stem= 0; R: stem= 1	0.045	0.036
4	CUB +she	0.39	0.564	0.030	-	2	L: north= 0; R: north= 1	0.037	0.022
5	CUB +she	0.45	0.631	0.082	-	6	-	0.059	0.014
6	CUB +she	0.52	0.662	0.020	-	8	L: abroad= 0; R: abroad= 1	0.053	0.044
7	CUB +she	0.79	0.695	0.016	-	2	L: gender= 0; R: gender= 1	0.046	0.036
8	CUB	0.34	0.610	0.000	-	-	-	0.027	0.027
9	CUB +she	0.44	0.516	0.034	-	2	-	0.071	0.055
12	CUB +she	0.63	0.669	0.033	-	8	L: gender= 0; R: gender= 1	0.043	0.028
13	CUB +she	0.41	0.643	0.048	-	2	-	0.076	0.052
14	CUB +she	0.77	0.718	0.016	-	2	L: north= 0; R: north= 1	0.042	0.034
15	CUB +she	0.73	0.649	0.079	-	6	-	0.065	0.028
24	CUB +she	0.62	0.708	0.072	-	8	-	0.088	0.051
25	CUB +she	0.60	0.638	0.025	-	5	-	0.031	0.025
28	CUB +she	0.63	0.739	0.026	-	4	-	0.033	0.021
29	CUB +she	0.75	0.706	0.058	-	6	-	0.064	0.044

Further literature on tree methods

Loh, W.Y., (2014). Fifty years of classification and regression trees, *International Statistical Review*, **82**(3): 329–348.

Sciandra, M., Plaia, A., Capursi, V. (2017). Classification trees for multivariate ordinal response: an application to student evaluation teaching. *Quality and Quantity*, **51**:641–655.

Tutz, G., Berger, M. (2021). Tree-structured scale effects in binary and ordinal regression. *Stat Comput*, 31, 17.

. . . . . .

Tree ensembles: The state of the art for ordered variables


Garge, N.R., Bobashev, G. & Eggleston, B. (2013). Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC Bioinformatics* 14, 125.


Buri M, Hothorn T. (2020). Model-based random forests for ordinal regression. *Int J Biostat.*, doi: 10.1515/ijb-2019-0063.


Janitza, S., Tutz, G., Boulesteix, A.L. (2016). Random Forests for Ordinal Responses: Prediction and Variable Selection. *Computational Statistics and Data Analysis* **96**,57–73


Wright, M.N., Ziegler A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, **77**(1), 1-17.


Hornung, R. (2020). Ordinal Forests. *Journal of Classification*. **37**: 4–17.


Simone, R., Tutz G. (2020). Hybrid random forests for ordinal data, In: *Book of Short Papers SIS 2020*, Pearson, Eds. A. Pollice, N. Salvati, and F. Schirripa Spagnolo, ISBN: 9788891910776, pp. 1171–1176.


Tutz, G. (2022). Ordinal Trees and Random Forests: Score-Free Recursive Partitioning and Improved Ensembles. *Journal of Classification* **39**:241–263.


. . . . . .


Simone, R. (2023). CUB random forests to assess the impact of uncertainty specification on prediction of ordinal scores (*preprint*)

Thank you for the attention!