Tree-structured scale effects in binary and ordinal regression

Tutz, Gerhard; Berger, Moritz

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Tutz, G., & Berger, M. (2021). Tree-structured scale effects in binary and ordinal regression. *Statistics and Computing*, *31*(2), 1-12. https://doi.org/10.1007/s11222-020-09992-0

Statistics and Computing (2021) 31:17
https://doi.org/10.1007/s11222-020-09992-0

Tree-structured scale effects in binary and ordinal regression

Gerhard Tutz[1] · **Moritz Berger**[2]

Received: 27 February 2020 / Accepted: 24 December 2020 / Published online: 9 February 2021

Abstract
In binary and ordinal regression one can distinguish between a location component and a scaling component. While the former determines the location within the range of the response categories, the scaling indicates variance heterogeneity. In particular since it has been demonstrated that misleading effects can occur if one ignores the presence of a scaling component, it is important to account for potential scaling effects in the regression model, which is not possible in available recursive partitioning methods. The proposed recursive partitioning method yields two trees: one for the location and one for the scaling. They show in a simple interpretable way how variables interact to determine the binary or ordinal response. The developed algorithm controls for the global significance level and automatically selects the variables that have an impact on the response. The modeling approach is illustrated by several real-world applications.

Keywords Recursive partitioning · Tree-structured modeling · Location-scale model · Heterogeneity of variances · Ordinal responses

1 Introduction

Tree-based models are strong nonparametric tools that allow to investigate interaction effects of covariates on responses. The basic concept is very simple: By binary recursive partitioning the predictor space is partitioned into a set of rectangles and on each rectangle a simple model (for example a constant) is fitted. The most popular versions are CART (Breiman et al. 1984), which is an abbreviation for classification and regression trees, and conditional inference trees, abbreviated by CTREE (Hothorn et al. 2006). Introductions and overviews were given, among others, by Loh (2014) and Strobl et al. (2009). Recursive partitioning methods, or simply trees, have several advantages: (i) they can be used in high-dimensional settings because they provide automatic variable selection, (ii) they have a built-in interaction detector, and (iii) they are easy to interpret and visualize. Besides classical regression trees for metrically scaled response variables, also versions for binary and ordinal responses are available, see Piccarreta (2008), Archer (2010) and Galimberti et al. (2012).

The objective of the present paper is to introduce trees in regression structures with ordinal responses that include scale effects, which are needed if unobserved heterogeneity of variances is present. The modeling of scale effects in ordinal regression was already considered by McCullagh (1980), who introduced the so-called *location-scale model* and gave a simple example with one binary covariate dealing with the quality of right eye vision for men and women. The location-scale model was considered and extended, among others, by Cox (1995) and Tutz and Berger (2017); Ishwaran and Gatsonis (2000) investigated the link to ROC analysis; Hedeker et al. (2008), Hedeker et al. (2009, 2012) showed how to use it in the case of repeated ordinal measurements.

Scale effects are also found in binary data. Their potential impact found much attention since Allison (1999) demonstrated that comparisons of binary model coefficients across groups can be misleading if one has underlying heterogeneity of residual variances. The problem has been investigated in various papers since then, see Williams (2009), Mood (2010), Karlson et al. (2012), Breen et al. (2014) and Rohwer (2015). One strategy to account for heterogeneity is to use McCullagh's location-scale model, which in the social sciences is also known as the heterogeneous choice or heteroskedastic logit model (Alvarez and Brehm 1995; Williams 2009). It

✉ Gerhard Tutz
tutz@stat.uni-muenchen.de

[1] Ludwig-Maximilians-Universität München, Akademiestrasse 1, 80799 Munich, Germany

[2] Institut für Medizinische Biometrie, Informatik und Epidemiologie, Medizinische Fakultät, Universitätsklinikum Bonn, Bonn, Germany

is included in various program packages as Stata, Limdep, SAS, and R.

As a parametric model that uses linear predictors the location-scale model is rather restrictive. In particular interactions of higher order are hard to include and lower order interactions are restricted to linear interactions. Tree-based methods offer a nonparametric alternative to investigate the interaction structure and automatically select variables. Variable selection is important since typically it is not known which variables contribute to location and to scaling. Since there are two components in the model, location and scaling, classical recursive partitioning methods cannot be used. The method developed in the following is explicitly designed to account for these two components. Two separate trees are obtained, one for each component.

In Sect. 2 the basic approach is introduced and illustrated by an application. In Sect. 3 the proposed algorithm is given in detail. More applications are considered in Sect. 5. The paper concludes with a summary given in Sect. 6.

2 Trees with scale effects

In the following we first consider basic ordinal models and the problems that might occur if variance heterogeneity is ignored. Then, we introduce the tree-structured modeling approach that is proposed.

2.1 Proportional odds and location-scale model

A common way to derive ordinal regression models is to assume that a latent variable is behind the ordinal response Y. Let the latent regression model have the form

$$Y_i^* = \alpha_0 + \boldsymbol{x}_i^T \boldsymbol{\alpha} + \sigma \varepsilon_i, \quad i = 1, \ldots, n\,,$$

where Y_i^* is the latent variable, $\boldsymbol{x}_i$ is a vector of covariates, and σ is the standard deviation of the noise variable ε_i, which has symmetric distribution function $F(.)$. The essential concept is to consider the ordinal response as a categorized version of the latent variable with the link between the observable ordinal variable Y_i with k categories and the latent variable Y_i^* given by

$$Y_i = r \Leftrightarrow \theta_{r-1} < Y_i^* \leq \theta_r\,, \tag{1}$$

where $-\infty = \theta_0 < \theta_1 < \cdots < \theta_k = \infty$ are thresholds on the latent scale. Simple derivation yields that the response probabilities are given by

$$P(Y_i \leq r|\boldsymbol{x}_i) = F\left(\frac{\alpha_{0r} - \boldsymbol{x}_i^T \boldsymbol{\alpha}}{\sigma}\right),$$

where $\alpha_{0r} = \theta_r - \alpha_0$. However, the model parameters are not identifiable. An identifiable version is obtained by setting $\sigma = 1$ or, equivalently, using $\beta_{0r} = \alpha_{0r}/\sigma$, $\boldsymbol{\beta} = \boldsymbol{\alpha}/\sigma$, which yields the *cumulative model*

$$P(Y_i \leq r|\boldsymbol{x}_i) = F(\beta_{0r} - \boldsymbol{x}_i^T \boldsymbol{\beta})\,. \tag{2}$$

The most prominent member of the family of cumulative models is the *proportional odds model*, which uses the logistic distribution function $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$. It has the form

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \eta_{ir} = \beta_{0r} - \boldsymbol{x}_i^T \boldsymbol{\beta}\,. \tag{3}$$

The strength of model (3) is that the parameters have an easily accessible interpretation. Let $\gamma_r(\boldsymbol{x}_i) = P(Y_i > r|\boldsymbol{x}_i)/P(Y_i \leq r|\boldsymbol{x}_i)$ denote the cumulative odds for category r. Then, one can derive that the effect of the jth variable is given by

$$e^{\beta_j} = \frac{\gamma_r(x_{i1}, \ldots, x_{ij} + 1, \ldots, x_{ip})}{\gamma_r(x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})}\,, \tag{4}$$

which does not depend on r. That means that e^{β_j} represents the multiplicative change in cumulative odds if x_{ij} increases by one unit for each category. Of course, the interpretation holds only if the model holds or is at least a good approximation to the data generating model.

It has been shown that the cumulative model (2) can yield very misleading results if there is variance heterogeneity in the underlying continuous regression model. Allison (1999) considered an example with the binary response being the promotion to an associate professor from the assistant professor level. It turned out that the number of published articles had a much stronger effect for male researchers than for female researchers, which seems rather unfair. He demonstrated that this effect could be due to heterogeneous variances.

The effect of heterogeneous variances is easily seen. Let the latent regression model be given by $Y_i^* = \alpha_0 + \boldsymbol{x}_i^T \boldsymbol{\alpha} + \sigma_i \varepsilon_i$, where σ_i now depends on the specific observation i. In the simplest case one has $\sigma_i = z_i \gamma$, where z_i is an indicator variable, which takes the value one for group 1 (for example males) and the value zero for group 0 (for example females). Then, the simple cumulative model (2) is mis-specified. The derivation from the latent variable yields

$$\begin{aligned} &P(Y_i \leq r|\boldsymbol{x}_i) = F(\alpha_{0r}/\sigma - \boldsymbol{x}_i^T(\boldsymbol{\alpha}/\sigma)) \\ &\quad \text{for observations from group 1 and} \\ &P(Y_i \leq r|\boldsymbol{x}_i) = F(\alpha_{0r} - \boldsymbol{x}_i^T \boldsymbol{\alpha}) \\ &\quad \text{for observations from group 0}\,. \end{aligned} \tag{5}$$

Thus, effects of covariates differ between the groups. One has $\boldsymbol{\alpha}/\sigma$ in group 1 and $\boldsymbol{\alpha}$ in group 0. If, for example, $\sigma = 0.5$ the effect strength in group 1 is twice the effect strength in group 0. The dependence on the group is simply ignored if one sets $\sigma = 1$, which is typically assumed in categorical regression. It means that in both groups the same scaling is used, although different ones are needed, see also Williams (2009), Mood (2010).

This form of mis-specification can be avoided by explicit modeling of the heterogeneity of variances. Let the standard deviation be determined by $\sigma_i = \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})$, where $\boldsymbol{z}_i$ is an additional vector of covariates, then one obtains from assumption (1) the *location-scale model*

$$P(Y_i \leq r|\boldsymbol{x}_i, \boldsymbol{z}_i) = F\left(\frac{\beta_{0r} - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}\right), \tag{6}$$

which for the logistic distribution function yields

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i, \boldsymbol{z}_i)}{P(Y_i > r|\boldsymbol{x}_i, \boldsymbol{z}_i)}\right) = \eta_{ir} = \frac{\beta_{0r} - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}. \tag{7}$$

The model contains two terms in the predictor that specifies the impact of covariates. The first is the location term $\beta_{0r} + \boldsymbol{x}_i^T\boldsymbol{\beta}$, and the second is the variance or scaling term $\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})$, which derives from the "variance equation" $\sigma_i = \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})$. Importantly, if $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are distinct, the interpretation of the $\boldsymbol{x}$-variables is the same as in the proportional odds model. With $\gamma_r(\boldsymbol{x}_i, \boldsymbol{z}_i) = P(Y_i > r|\boldsymbol{x}_i, \boldsymbol{z}_i)/P(Y_i \leq r|\boldsymbol{x}_i, \boldsymbol{z}_i)$ denoting the cumulative odds for category r one obtains again the relation (4) and therefore an interpretation of parameters that does not depend on the category.

The location-scale model was introduced by McCullagh (1980) but is also known as *heterogeneous choice model* or heteroskedastic logit model (Alvarez and Brehm 1995). It should be noted that although the scaling component is typically motivated from variance heterogeneity it can also be seen as representing interactions or effect-modifying effects, see Rohwer (2015) and Tutz (2018). As Williams (2010) noted, it is also strongly related to the logistic response model with proportionality constraints proposed by Hauser and Andrew (2006) and extended by Fullerton and Xu (2012).

2.2 Tree-structured location-scale models

Recursive partitioning methods for ordinal responses have been proposed by Archer (2010), Galimberti et al. (2012) and are available in R packages. Also the conditional unbiased recursive partitioning framework as proposed by Hothorn et al. (2006) allows to fit trees for ordinal responses. However, all of these methods do not account for possible heterogeneity induced by variance.

The problem with modeling heterogeneity is that one has to fit two separate predictors, the location term and the variance term. In the traditional location-scale model (6) they are represented by the linear predictor $\beta_{0r} - \boldsymbol{x}_i^T\boldsymbol{\beta}$ and the variance term $\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})$, respectively. The tree proposed here also distinguishes between location and variance; for both components separate trees are fitted. It is crucial that the partitioning of location and variance terms has to be done in a coordinated way. Trees have to be grown by taking both components into account simultaneously.

In the following, we first sketch the basic algorithm, which will be given in more detail in Sect. 3. The basic concept is to replace the predictor $\eta_{ir} = (\beta_{0r} - \boldsymbol{x}_i^T\boldsymbol{\beta})/\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})$ of the location-scale model (6) by coordinated recursive partitioning terms.

Basic algorithm

Let us consider the building of a tree when starting at the root. We will first focus on metrically scaled and ordinal (including binary) covariates. In this case the partition of a node A into two subsets A_1 and A_2 has the form

$$A_1 = A \cap \{x_j \leq c\} \quad \text{and} \quad A_2 = A \cap \{x_j > c\},$$

with regard to threshold c on variable x_j.

First step

For each variable x_j and all corresponding thresholds c that can be built for this variable one investigates the following fits:

(a) Location term:
One fits the location-scale model with one split in the location term and predictor

$$\eta_{ir} = \beta_{0r} - \beta I(x_{ij} \leq c),$$

where $I(.)$ is the indicator function. Then, one obtains

$$\eta_{ir} = \beta_{0r} - \beta \quad \text{if } x_{ij} \leq c \quad \text{and}$$
$$\eta_{ir} = \beta_{0r} \quad \text{if } x_{ij} > c.$$

Alternatively, one can replace $I(.)$ by $I^*(.) = 2I(.) - 1$, which means one uses effect coding and replaces the $0-1$ dummy variable by the variable $I^*(.) = 1$ if $x_{ij} \leq c$ and $I^*(.) = -1$ otherwise. Accordingly, one obtains

$$\eta_{ir} = \beta_{0r} - \beta \quad \text{if } x_{ij} \leq c \quad \text{and}$$
$$\eta_{ir} = \beta_{0r} + \beta \quad \text{if } x_{ij} > c.$$

(b) Variance term:
One fits the location-scale model with one split in the variance term and predictor

$$\eta_{ir} = \frac{\beta_{0r}}{\exp(\gamma I(x_{ij} \leq c))}.$$

Then, one obtains

$$\eta_{ir} = \frac{\beta_{0r}}{\exp(\gamma)} \quad \text{if } x_{ij} \leq c \quad \text{and}$$
$$\eta_{ir} = \beta_{0r} \qquad \text{if } x_{ij} > c\,.$$

One chooses the best split according to an appropriate splitting criterion (for details, see Sect. 3) among all the fitted models from (a) and (b). Thus, in the first step one split is performed either in the location term or the variance term.

Later steps

In later steps the splitting is done in a similar way. Let $A_1^{\text{loc}}, \ldots, A_{m_{\text{loc}}}^{\text{loc}}$ denote the nodes (subsets of the predictor space) of the location term from the previous steps. Accordingly, let $A_1^{\text{sc}}, \ldots, A_{m_{\text{sc}}}^{\text{sc}}$ denote the nodes (subsets of the predictor space) of the variance term from the previous steps. Note that, all nodes are determined by a product of indicator functions. For example, if the splits were in the metric variables x_3 and x_7 a node may be determined by $I(\boldsymbol{x}_i \in A) = I(x_{i3} > 20)I(x_{i7} \leq 4)$.
One fits all the candidate models

(a) for the splitting of A_k^{loc}, $k = 1, \ldots, m_{\text{loc}}$ in the location term with predictors

$$\frac{\beta_{0r} - \sum_{s=1}^{m_{\text{loc}}} \beta_s I(\boldsymbol{x}_i \in A_s^{\text{loc}}) - \beta I(\boldsymbol{x}_i \in A_k^{\text{loc}}) I(x_{ij} \leq c)}{\exp\left(\sum_{\ell=1}^{m_{\text{sc}}} \gamma_\ell I(\boldsymbol{x}_i \in A_\ell^{\text{sc}})\right)}$$

to obtain the $(m_{\text{loc}} + 1)$th node in the location term with parameter estimate β,

(b) for the splitting of A_k^{sc}, $k = 1, \ldots, m_{\text{sc}}$ in the variance term with predictors

$$\frac{\beta_{0r} - \sum_{s=1}^{m_{\text{loc}}} \beta_s I(\boldsymbol{x}_i \in A_s^{\text{loc}})}{\exp(\sum_{\ell=1}^{m_{\text{sc}}} \gamma_\ell I(\boldsymbol{x}_i \in A_\ell^{\text{sc}}) + \gamma I(\boldsymbol{x}_i \in A_k^{\text{sc}}) I(x_{ij} \leq c))}\,.$$

to obtain the $(m_{\text{sc}} + 1)$th node in the variance term with parameter estimate γ.

One chooses the best split according to an appropriate splitting criterion among all the possible models from (a) and (b). Again, each step means an update of the location term or the variance term. After termination of the algorithm according to an appropriate stopping criterion, the final model consists of two trees: one for the location component and one for the scale component, with different partitions.

We refer to the concept as *tree-structured model building* to distinguish it from the *model-based* recursive partitioning models as considered by Zeileis et al. (2008). The basic idea of model-based recursive partitioning is to fit models in subspaces of the predictor space and then decide which partitioning explains the predictor–response relationships best. Of course elaborated methods are needed to ensure that the splits represent relevant information, for example, by using appropriate tests, see Zeileis et al. (2008). Although in principle this approach could also be used in the location-scale framework the obtained tree would not separate between the two types of influential terms. The main difference between tree-structured modeling and model-based recursive partitioning is that tree-structured model building means that *the predictor structure is determined by trees*, whereas model-based approaches do not structure the predictor but fit the whole model in subspaces. Tree-structured modeling yields separate trees for the two influential terms: one tree for the location and one tree for the variance heterogeneity. Thus, it is easily seen which variables contribute to which component. Tree structures in the predictor have been considered before, but in a quite different context; Berger and Tutz (2017) and Tutz and Berger (2018) considered trees to model the effect of categorical predictors on the response if the predictors have a very large number of categories.

Before considering an illustrative example we briefly consider the interpretation of parameters. Let $A_1^{\text{loc}}, \ldots, A_{m_{\text{loc}}}^{\text{loc}}$ denote the end nodes of the location term, and $A_1^{\text{sc}}, \ldots, A_{m_{\text{sc}}}^{\text{sc}}$ denote the end nodes of the variance term. Then, one has the predictor

$$\eta_{ir} = \frac{\beta_{0r} - \sum_{s=1}^{m_{\text{loc}}} \beta_s I(\boldsymbol{x}_i \in A_s^{\text{loc}})}{\exp(\sum_{\ell=1}^{m_{\text{sc}}} \gamma_\ell I(\boldsymbol{x}_i \in A_\ell^{\text{sc}}))}, r = 1, \ldots, k-1\,.$$

The interpretation is similar to the interpretation of parameters in the location-scale model, the β-parameters indicate the location and the γ-parameters variance heterogeneity. For illustration let us consider extreme cases.

- If $\beta_s \to -\infty$, one obtains for $\boldsymbol{x}_i \in A_s^{\text{loc}}$ (fixed variance component) the probabilities $P(Y_i = 1|\boldsymbol{x}_i) = 1$, and $P(Y_i = 2|\boldsymbol{x}_i) = \ldots P(Y_i = k|\boldsymbol{x}_i) = 0$. If $\beta_s \to \infty$, one obtains for $\boldsymbol{x}_i \in A_s^{\text{loc}}$ the probabilities $P(Y_i = k|\boldsymbol{x}_i) = 1$, and $P(Y_i = 1|\boldsymbol{x}_i) = \ldots P(Y_i = k-1|\boldsymbol{x}_i) = 0$. That means the size of β_s indicates the preference for high categories.
- If $\gamma_\ell \to \infty$, one obtains for $\boldsymbol{x}_i \in A_\ell^{\text{loc}}$ (fixed location component) the probabilities $P(Y_i = 1|\boldsymbol{x}_i) = P(Y_i = k|\boldsymbol{x}_i) = 0.5$, which means maximal heterogeneity with all responses in the extreme categories.

Nominal covariates

For a categorical covariate with K unordered categories $x_j \in \{1, \ldots, K\}$, the partition of a node A has the form $A \cap S$ and $A \cap \bar{S}$, where S and $\bar{S}$ are disjoint, non-empty subsets $S \subset \{1, \ldots, K\}$ and $\bar{S} = \{1, \ldots, K\} \setminus S$. Thus, one has $2^{K-1} - 1$ possible splits. For large K the number of candidate splits is excessive, it increases computational complexity and restricts the possible number of categories that can be sensibly used.

For continuous and binary responses it has been shown that ordering the categories by increasing means of the response and treating these ordered categories as ordinal also leads to the optimal splits (Fisher 1958; Breiman et al. 1984). This reduces computational complexity because only $K - 1$ splits have to be considered.

For categorical responses, Wright and König (2019) proposed a sorting algorithm for ordering the categories, which is based on a approximate solution by Coppersmith et al. (1999). For each categorical covariate, the basic steps of the algorithm are the following:

1. Compute the probability matrix $\boldsymbol{P} \in \mathbb{R}^{Kxk}$, where the rows contain the relative class frequencies conditionally on the covariate categories.
2. Compute the covariance matrix $\boldsymbol{S} \in \mathbb{R}^{KxK}$ from $\boldsymbol{P}$, weighted by the absolute frequency of the covariate categories.
3. Sort the covariate categories by the scores of the first principle component of $\boldsymbol{S}$.

Particularly, Wright and König (2019) show that it is sufficient to order the categories a priori, that is, once on the entire data before the analysis (but not in every split during tree building). This approach results in faster computation, does not suffer from a category limit problem and has the advantage that categories not present in a node can still be assigned to a child node. In our R program we make use of the sorting algorithm by Wright and König (2019) prior to tree building and subsequently treat categorical variables as ordinal.

2.3 Illustrative example

Confidence data

We consider data from the general social survey of social science, in short ALLBUS, a study by the German institute GESIS. The data are available from http://www.gesis.org/allbus. Our analysis is based on a subset containing 2935 respondents of the ALLBUS in 2012. The response is the confidence in the federal government measured on a symmetric scale from 1 (no confidence at all/excessive distrust) to 7 (excessive confidence). As explanatory variables we consider the gender (0: male, 1: female), the income in thousands of Euros, the age in decades (centered at 50) and the self reported interest in politics from 1 (very strong interest) to 5 (no interest at all).

Figure 1 shows the tree obtained for the location term and Fig. 2 the tree for the variance term. It is seen that the main drivers of confidence are interest in politics and age. Among respondents that have strong interest in political issues (interest = 5) those above 40 years of age have weak confidence (node 5), whereas those below 40 years tend to prefer higher categories (node 4). Among respondents that are less interested in politics, in particular young people (age lesser than 25) and older people (age above 74) show a strong tendency to choose high confidence categories ($\hat{\beta}_s = 0.951$ and $\hat{\beta}_s = 0.824$). From the variance tree it is seen that males with low income (node 4; $\hat{\gamma}_\ell = 0.214$) are the most heterogeneous groups with comparatively large variance, whereas females form the most homogeneous groups.

3 The algorithm in detail

In all tree-based methods, one has to decide in particular how to split and how to determine the size of the trees. In traditional approaches, one typically grows large trees and prunes them to an adequate size afterward, see Breiman et al. (1984) and Ripley (1996). An alternative strategy, which was propagated within the conditional unbiased recursive partitioning framework (Hothorn et al. 2006), is to directly control the size of the trees by early stopping. We also use this approach and control the significance of splits by using tests for cumulative regression models.

Let us consider again the construction of the first split. A split in the location term with regard to the jth variable yields the model with predictor

$$\eta_{ir} = \beta_{0r} - \beta_j I(x_{ij} \leq c_j),$$

and a split in the variance term with regard to the jth variable yields the model with predictor

$$\eta_{ir} = \frac{\beta_{0r}}{\exp(\gamma_j I(x_{ij} \leq c_j))}.$$

To test for the best split among all the covariates, the set of possible split points and the two components (location or variance) one examines all the null hypotheses $H_0 : \beta_j = 0$ and $H_0 : \gamma_j = 0$ and selects that split as the optimal one that has the smallest p value. As test statistic, we use the LR test statistic. Computing the LR test statistic requires fitting of both models, the full model and the restricted model under H_0. We nevertheless prefer the LR statistic because it

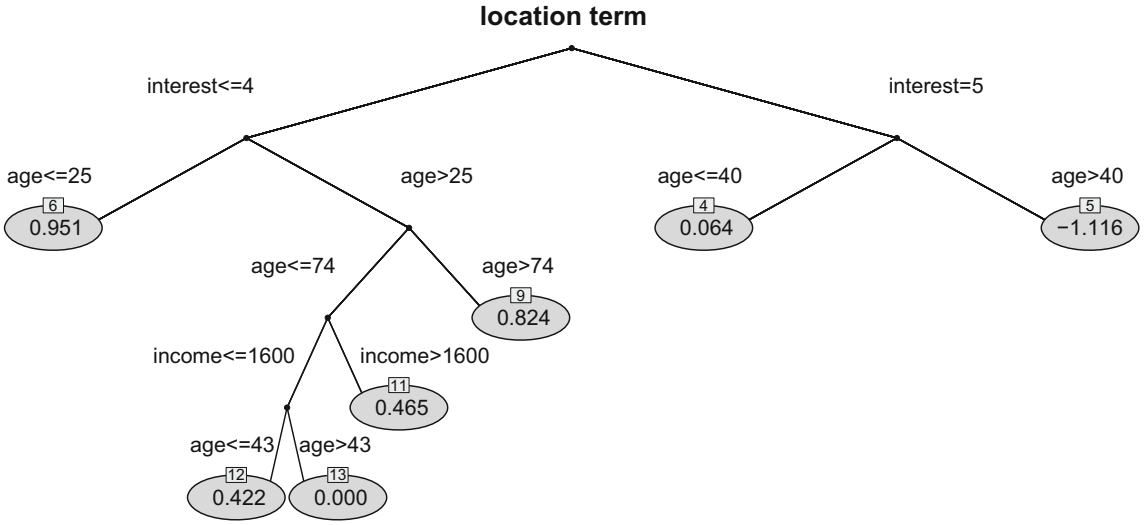

Fig. 1 Tree for location term of confidence data. The parameter estimates $\hat{\beta}_s$ are given in the terminal nodes

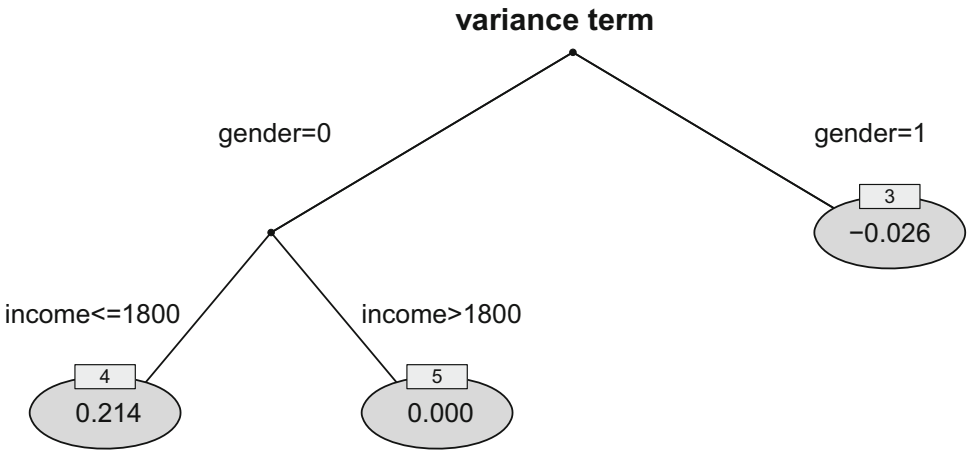

Fig. 2 Tree for variance term of confidence data. The parameter estimates $\hat{\gamma}_\ell$ are given in the terminal nodes

corresponds to selecting the model with minimal deviance. This criterion is also equivalent to minimizing the entropy, which belongs to the family of impurity measures.

To decide whether the selected split should be performed, we apply a concept based on maximally selected statistics. The basic idea is to investigate the dependence of the ordinal response and the selected variable at a global level that takes the number of splits into account. For one fixed component and variable j, one simultaneously considers all LR test statistics T_{jc_j}, where c_j are from the set of possible split points, and computes the maximal value statistic $T_j = \max_{c_j} T_{jc_j}$. The p value that can be obtained by the distribution of T_j provides a measure for the relevance of variable j. The result is not influenced by the number of split points; therefore, the method explicitly accounts for the involved multiple testing problem; for similar approaches, which inspired the proposed method, see Hothorn and Lausen (2003), Shih (2004), Shih and Tsai (2004) and Strobl et al. (2007). As the distribution of T_j in general is unknown we use a permutation test to obtain a decision on the null hypothesis. The distribution of T_j is determined by computing the maximal value statistics based on random permutations of variable j. A random permutation of variable j breaks the relation of the covariate and the response in the original data. By computing the maximal value statistics for a large number of permutations one obtains an approximation of the distribution under the null hypothesis and the corresponding p value. Importantly, to determine the p value with sufficient accuracy, the number of permutations should increase with the number of covariates.

In all later steps the basic procedure is the same; one searches for the statistic with the maximal value trying all combinations of variables and split points in both components. For the components that already have been split (location, variance or both) one starts from already built nodes. Given overall significance level α the significance level for the permutation test that tests splits in one variable is

chosen by $\alpha/2p$, where p denotes the number of covariates that are available in the two components.

Altogether, the following steps are carried out during the fitting procedure:

1. *(Initial model)* Fit the model with category-specific intercepts only, yielding the estimates $\hat{\beta}_{01}, \ldots, \hat{\beta}_{0,k-1}$.
2. *(Tree building)*
   (a) For all explanatory variables x_j, $j = 1, \ldots, p$, fit all the candidate models with one additional split in one of the already built nodes in both components.
   (b) Select the best model using the p values of the LR test statistics.
   (c) Carry out the permutation test for the selected node (defined by a combination of variable, split point and component) using the maximal value statistic with significance level $\alpha/2p$. If significant, fit the selected model and continue with Step 2(a), else continue with Step 3.
3. *(Selected model)* Fit the final model with components $\hat{\beta}_{0r}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$.

The final model consists of one or two separate trees: one referring to the location component and one referring to the variance component. In general, the trees will be different but can also yield the same partitioning. It should be noted that in contrast to the way trees are grown in traditional recursive partitioning all parameter estimates change if an additional split is performed.

Prediction for new observations

For a (new) observation with covariates $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{z}}_i$ one obtains predictions of the cumulative odds by identifying the corresponding terminal nodes of the two trees and computing

$$\hat{\eta}_{ir} = \frac{\hat{\beta}_{0r} - \sum_{s=1}^{m_{\text{loc}}} \hat{\beta}_s I(\tilde{\boldsymbol{x}}_i \in A_s^{\text{loc}})}{\exp\left(\sum_{\ell=1}^{m_{\text{sc}}} \hat{\gamma}_\ell I(\tilde{\boldsymbol{z}}_i \in A_\ell^{\text{sc}})\right)},$$

and

$$\frac{P(Y_i \leq r | \boldsymbol{x}_i, \boldsymbol{z}_i)}{P(Y_i > r | \boldsymbol{x}_i, \boldsymbol{z}_i)} = \exp\left(\hat{\eta}_{ir}\right), \; r = 1, \ldots, k-1\,.$$

4 Simulation study

In this section, we present the results of numerical experiments to investigate the performance of the proposed modeling approach. The primary aim of the study is to analyze the ability of the tree-structured algorithm to correctly detect the informative covariates in both the location term and the variance term.

4.1 Experimental design

In all simulations scenarios the ordinal responses $Y_i \in \{1, \ldots, 5\}$, $i = 1, \ldots, n$, were simulated from the location-scale model (6) with differing specifications of the predictor functions η_{ir}. We generated datasets with $n \in \{500, 1000\}$ observations (1000 replications each), and included two standard normally distributed covariates, $x_1, x_2 \sim N(0, 1)$, two binary covariates, $x_3, x_4 \sim B(1, 0.5)$ and two nominal covariates with four categories $x_5, x_6 \sim M(1, 0.25)$. The category-specific intercepts were set to $\beta_{0r} \in \{-0.25, -0.08, 0.08, 0.25\}$. All permutation tests were based on 1200 permutations with overall significance level $\alpha = 0.05$.

Evaluation criteria

In order to evaluate the performance of the algorithm we computed true positive rates (TPR) and false positive rates (FPR) for the location term and variance term, respectively. Let δ_j^{loc} and δ_j^{sc} $j = 1, \ldots, 4$, be indicators with $\delta_j^{\text{loc}} = 1$ if covariate x_j is influential in the location term and $\delta_j^{\text{sc}} = 1$ if covariate x_j is influential in the variance term. Otherwise, the two indicators are equal to zero. Then with indicator function $I(\cdot)$, the used performance measures are:

– True positive rate in the location term:

$$\text{TPR}^{\text{loc}} = \frac{1}{\#\{j : \delta_j^{\text{loc}} = 1\}} \sum_{j:\delta_j^{\text{loc}}=1} I(\hat{\delta}_j^{\text{loc}} = 1)$$

– True positive rate in the variance term:

$$\text{TPR}^{\text{sc}} = \frac{1}{\#\{j : \delta_j^{\text{sc}} = 1\}} \sum_{j:\delta_j^{\text{sc}}=1} I(\hat{\delta}_j^{\text{sc}} = 1)$$

– False positive rate in the location term:

$$\text{FPR}^{\text{loc}} = \frac{1}{\#\{j : \delta_j^{\text{loc}} = 0\}} \sum_{j:\delta_j^{\text{loc}}=0} I(\hat{\delta}_j^{\text{loc}} = 1)$$

– False positive rate in the variance term:

$$\text{FPR}^{\text{sc}} = \frac{1}{\#\{j : \delta_j^{\text{sc}} = 0\}} \sum_{j:\delta_j^{\text{sc}}=0} I(\hat{\delta}_j^{\text{sc}} = 1)$$

Simulation scenarios

We consider four simulation scenarios with the following true underlying predictor functions. In each case the influential terms correspond to trees with three terminal nodes.

– Scenario 1 without informative variables:

$$\eta_{ir} = \beta_{0r},\ r = 1, \ldots, 4\,.$$

– Scenario 2 with informative variables in the location term only:

$$\eta_{ir} = \beta_{0r} + \beta I(\{x_1 > 0\}) - 2\,\beta I(\{x_1 > 0\} \cap \{x_3 = 0\})\,,$$
$$\beta \in \{0.4, 0.6, 0.8\}\,.$$

– Scenario 3 with informative variables in the variance term only:

$$\eta_{ir} = \frac{\beta_{0r}}{\gamma I(\{x_2 > 0\}) - 2\,\gamma I(\{x_2 > 0\} \cap \{x_6 \in \{1, 3\}\})}\,,$$
$$\gamma \in \{0.5, 0.75, 1\}\,.$$

– Scenario 4 with informative variables in both terms:

$$\eta_{ir} = \frac{\beta_{0r} + \beta I(\{x_1 > 0\}) - 2\,\beta I(\{x_1 > 0\} \cap \{x_3 = 0\})}{\gamma I(\{x_2 > 0\}) - 2\,\gamma I(\{x_2 > 0\} \cap \{x_6 \in \{1, 3\}\})}\,,$$
$$\beta \in \{0.4, 0.6, 0.8\}, \quad \gamma \in \{0.5, 0.75, 1\}\,.$$

4.2 Results

Table 1 summarizes the results of the four simulation scenarios. Each value in the table corresponds to the average detection rate over 1000 replications. It is seen that the TPR (fourth and fifth column in Table 1) highly depend on the sample size and the true effect size. While for $n = 500$ and small effect size ($\beta = 0.4$ and/or $\gamma = 0.5$) the algorithm is not very efficient in detecting the influential covariates, the detection works quite perfect in the settings with $n = 1000$ and strong effect size ($\beta = 0.8$ and/or $\gamma = 1$). In the latter cases the TPR are all higher than 0.96. The results of scenario 4 (where different covariates are influential in the two components) further show that the procedure is well able to separate between the two types of influential terms, as the TPR are widely comparable to those in scenario 2 and scenario 3.

Regarding the FPR (sixth and seventh column in Table 1) the results demonstrate that the algorithm hardly includes one of the non-influential covariates. In scenario 1 without any influential covariates the procedure is most restrictive. Importantly, the FPR are below the overall significance level of $\alpha = 0.05$ in both terms throughout all settings, even with strong effects of the informative variables.

Table 1 Results of the simulation study

	β	γ	TPR^{loc}	TPR^{sc}	FPR^{loc}	FPR^{sc}
n = 500						
1	–	–	–	–	0.008	0.007
2	0.4	–	0.156	–	0.009	0.006
	0.6	–	0.465	–	0.018	0.012
	0.8	–	0.788	–	0.018	0.020
3	–	0.5	–	0.074	0.008	0.009
	–	0.75	–	0.320	0.011	0.008
	–	1	–	0.717	0.020	0.011
4	0.4	0.5	0.168	0.067	0.012	0.009
	0.6	0.5	0.522	0.054	0.016	0.012
	0.8	0.5	0.829	0.057	0.022	0.024
	0.4	0.75	0.203	0.254	0.013	0.010
	0.4	1	0.295	0.549	0.021	0.009
n = 1000						
1	–	–	–	–	0.007	0.007
2	0.4	–	0.437	–	0.015	0.009
	0.6	–	0.865	–	0.021	0.017
	0.8	–	0.989	–	0.019	0.024
3	–	0.5	–	0.263	0.010	0.009
	–	0.75	–	0.761	0.016	0.013
	–	1	–	0.983	0.021	0.022
4	0.4	0.5	0.476	0.240	0.015	0.013
	0.6	0.5	0.904	0.241	0.024	0.027
	0.8	0.5	0.993	0.264	0.027	0.032
	0.4	0.75	0.623	0.686	0.025	0.021
	0.4	1	0.841	0.964	0.041	0.029

TPR and FPR in the location term and in the variance term for $n = 500$ (upper panel) and $n = 1000$ (lower panel) averaged over 1000 replications, respectively. Note that the algorithm showed fitting problems in 0.3% (scenario 2), 0.2% (scenario 3) and 1.9% (scenario 4) of the replications, because of the ordering constraint on the intercepts in the cumulative model

5 Further applications

Biochemists data

Let use consider the application used by Allison (1999) when investigating the problem if effects of variables differ over gender groups. The dataset, which has also been used by Long et al. (1993) and Williams (2009), investigates the careers of 301 male and 177 female biochemists (the following description is adapted from Allison, 1999). Binary regression is used to predict the probability of promotion to associate professor from the assistant professor level (1: no promotion, 2: promotion). The variables in the model are the number of years since the beginning of the assistant professorship (years), undergraduate selectivity as a measure of the selectivity of the colleges where scientists received their

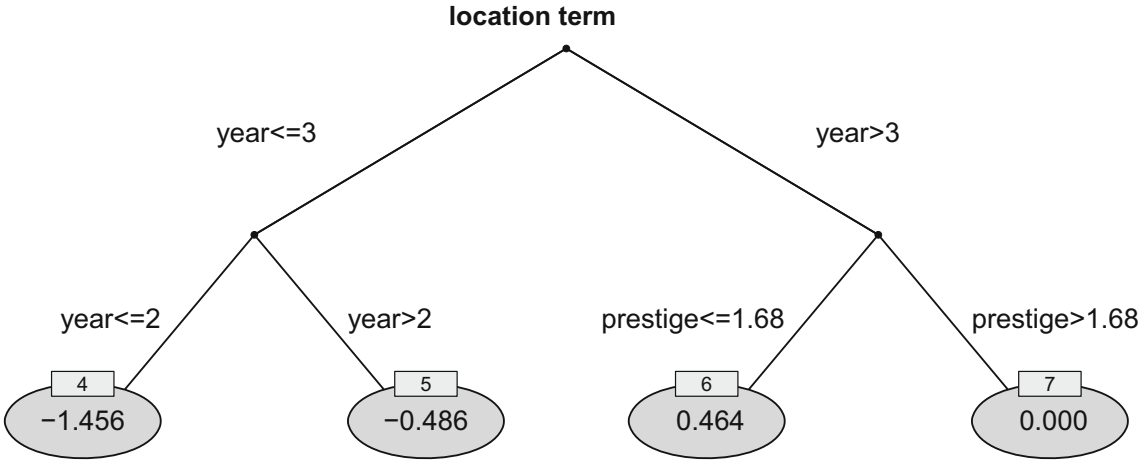

Fig. 3 Tree for location term of biochemists example. The parameter estimates $\hat{\beta}_s$ are given in the terminal nodes

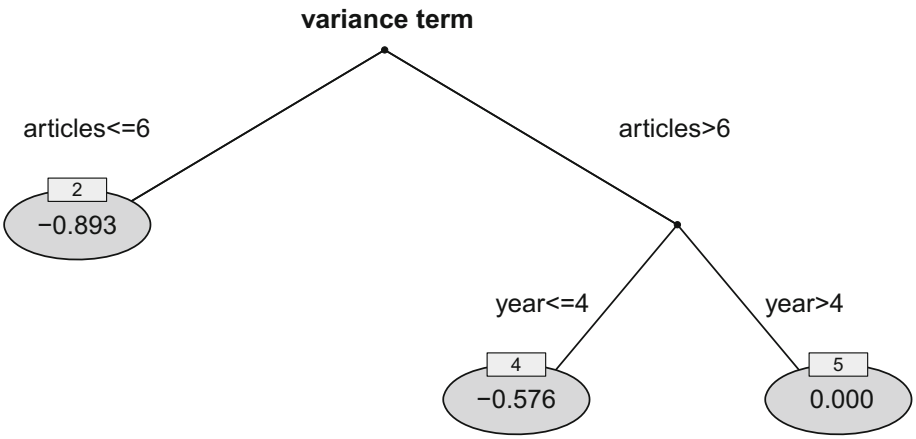

Fig. 4 Tree for variance term of biochemists example. The parameter estimates $\hat{\gamma}_\ell$ are given in the terminal nodes

bachelor's degrees (select), the number of articles (articles) representing the cumulative number of articles published by the end of each person year, and job prestige (prestige) measuring the prestige of the department in which scientists were employed. Figures 3 and 4 show the fitted trees for location and variance, respectively.

While Allison (1999) focused on gender as a relevant variable in the variance term, it is seen from the trees that gender does not seem to be very influential; neither in the location term nor in the variance term gender is present. A similar result was obtained by Williams (2010). When he used a stepwise forward strategy to select variables in the parametric location-scale model, the only variable that entered the variance equation was the number of articles. He also made a plausible argument for this by stating that "there may be little residual variability among biochemists with few articles (with most of them being denied tenure) but there may be much more variability among biochemists with more articles (having many articles may be a necessary but not sufficient condition for tenure)."

It is seen from the trees that the chances of a promotion to associate professor are best for biochemists who have spent at least three years at a department with not the highest prestige (node 6). Applicants with articles ≤ 6 or articles > 6 in combination with year ≤ 4 seem to form the most homogeneous groups.

To evaluate the issue of unfairness further we fitted trees when only the covariates gender and number of articles are included in the analysis. The corresponding trees are given in Fig. 5. It is seen that only the number of articles was found to have an impact on location as well as on variance. There is no indication that gender plays a crucial role for the promotion to associate professor.

Retinopathy data

In a 6-year follow-up study on diabetes and retinopathy status reported by Bender and Grouven (1998) the interesting question was how the retinopathy status is associated with risk factors. The considered risk factors were smoking (SM = 1: smoker, SM = 0: non-smoker), diabetes duration (DIAB)

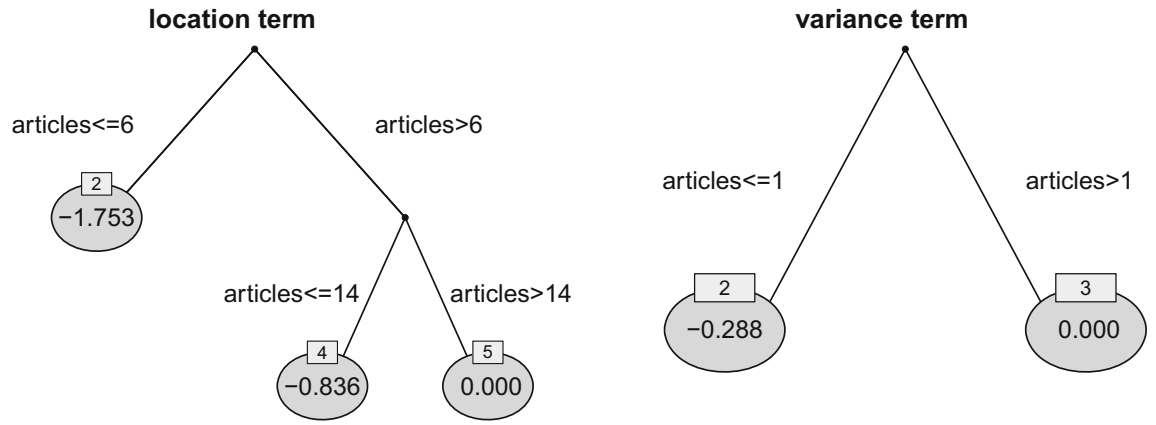

Fig. 5 Tree for location (left) and variance (right) of biochemists example with only gender and articles included. The parameter estimates $\hat{\beta}_s$ and $\hat{\gamma}_\ell$ are given in the terminal nodes, respectively

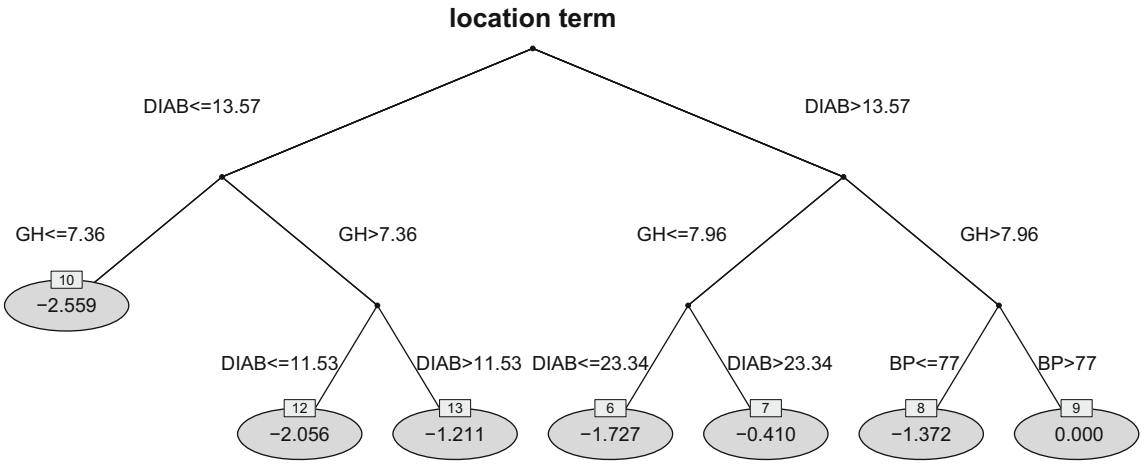

Fig. 6 Tree for location term of retinopathy data. The parameter estimates $\hat{\beta}_s$ are given in the terminal nodes

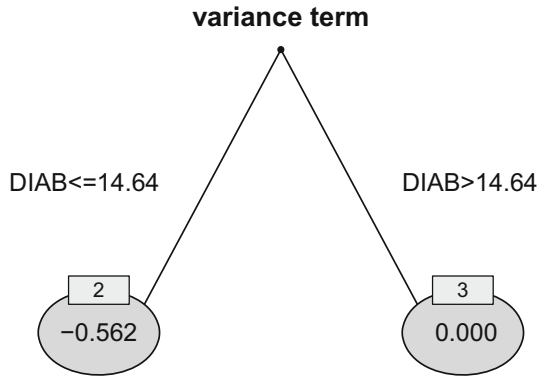

Fig. 7 Tree for variance term of retinopathy data. The parameter estimates $\hat{\gamma}_\ell$ are given in the terminal nodes

measured in years, glycosylated hemoglobin (GH), which is measured in percent, and diastolic blood pressure (BP) measured in mmHg. The response variable retinopathy status has three categories (1: no retinopathy; 2: nonproliferative retinopathy; 3: advanced retinopathy or blind).

It is seen from Fig. 6 that in particular the duration of diabetes is influential followed by glycosylated hemoglobin. The lowest risk is found in node 10 (DIAB $\leq$ 13.57, GH $\leq$ 7.36). Even if GH $>$ 7.36 but DIAB $\leq$ 11.53, the risk is still very low. The highest risks are found for long duration of diabetes DIAB $\leq$ 23.34 in combination with low values of glycosylated hemoglobin GH $\leq$ 7.96 (node 7) and in node 9, which combines long diabetes duration and high values of glycosylated hemoglobin and diastolic blood pressure. Figure 7 shows that patients with longer duration of diabetes are more homogeneous (sharing higher risk) than patients with lower values of diabetes duration.

Predictive performance

Finally, we compared the prediction accuracy of the tree-structured model to a single CTREE (Hothorn et al. 2006) in the three applications. For this, we repeatedly (100 replications) fitted the two models on subsamples without replacement containing 2/3 of the original dataset and com-

puted the ranked probability score from the remaining test datasets (i.e., from 1/3 of the original data). The ranked probability score is particularly appropriate for the evaluation of probability forecasts of ordinal variables (Murphy 1970).

For the confidence data we observed the values (mean (range)), 3.144 (3.095–3.196) when fitting the tree-structured model and 3.147 (3.101–3.198) when fitting a CTREE. For the biochemists data we observed the values 0.881 (0.845–0.910) and 0.882 (0.853–0.921) including all five covariates, and for the retinopathy data we obtained 1.484 (1.417–1.541) and 1.488 (1.384–1.581).

The results indicate that there is only minor improvement in prediction when using the tree-structured model, which fits the location-scale model, compared to a single tree. Our proposed method mainly serves as an explanatory tool showing which variables influence the location, and which variables influence the variance of the ordinal responses. If the objective is the best prediction, it is advisable to use random forest methods as proposed, for example, by Janitza et al. (2016) and Hornung (2020).

6 Summary and concluding remarks

Let us summarize the strengths of the proposed tree method.

- One obtains two trees: one for the location and one for the variance. Thus, it is clearly seen which variables have an impact on which component.
- The obtained trees have a simple interpretation showing which combinations of variables determine the preference of categories, and which sub-populations form more homogeneous or heterogeneous groups.
- By fitting a scale (or variance) component the method avoids misleading effects that may occur if one ignores potential variance heterogeneity.
- As in all tree-based methods interactions are explicitly modeled and there is a built-in variable selection procedure.

The presented algorithm is constructed such that only variables for which a significant effect can be detected are included. By controlling for the overall significance level the inclusion of irrelevant variables is avoided. These properties of the procedure are demonstrated in the simulation study. It has the effect that the procedure tends to include relatively few variables, in particular if many variables are available. However, the method can also be used in an exploratory way. If one uses a significance level distinctly larger than .05, one obtains much larger trees, which might hint at further possible interaction effects. Nevertheless, we think it is essential to control for the significance level, which gets lost in many procedures, especially if one first fits trees and then starts pruning as in conventional trees.

An R implementation of the proposed tree-structured model including an auxiliary function to plot the trees, as well as exemplary code to reproduce the illustrative example, is available from GitHub (https://github.com/jmober/LocationScaleTree).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Allison, P.D.: Comparing logit and probit coefficients across groups. Sociol. Methods Res. **28**(2), 186–208 (1999)

Alvarez, R.M., Brehm, J.: American ambivalence towards abortion policy: development of a heteroskedastic probit model of competing values. Am. J. Polit. Sci. **39**, 1055–1079 (1995)

Archer, K.J.: rpartordinal: an R package for deriving a classification tree for predicting an ordinal response. J. Stat. Softw. **34**, 7 (2010)

Bender, R., Grouven, U.: Using binary logistic regression models for ordinal data with non-proportional odds. J. Clin. Epidemiol. **51**, 809–816 (1998)

Berger, M., Tutz, G.: Tree-structured clustering in fixed effects models. J. Comput. Graph. Stat. **27**(2), 380–392 (2017). https://doi.org/10.1080/02664763.2017.1383370

Breen, R., Holm, A., Karlson, K.B.: Correlations and nonlinear probability models. Sociol. Methods Res. **43**(4), 571–605 (2014)

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, J.C.: Classification and Regression Trees. Wadsworth, Monterey (1984)

Coppersmith, D., Hong, S.J., Hosking, J.R.: Partitioning nominal attributes in decision trees. Data Min. Knowl. Discov. **3**(2), 197–217 (1999)

Cox, C.: Location-scale cumulative odds models for ordinal data: a generalized non-linear model approach. Stat. Med. **14**(11), 1191–1203 (1995)

Fisher, W.D.: On grouping for maximum homogeneity. J. Am. Stat. Assoc. **53**(284), 789–798 (1958)

Fullerton, A.S., Xu, J.: The proportional odds with partial proportionality constraints model for ordinal response variables. Soc. Sci. Res. **41**(1), 182–198 (2012)

Galimberti, G., Soffritti, G., Maso, M.D., et al.: Classification trees for ordinal responses in R: the rpartscore package. J. Stat. Softw. **47**(i10), 1–25 (2012)

Hauser, R.M., Andrew, M.: 1. Another look at the stratification of educational transitions: the logistic response model with partial proportionality constraints. Sociol. Methodol. **36**(1), 1–26 (2006)

Hedeker, D., Mermelstein, R.J., Demirtas, H.: An application of a mixed-effects location scale model for analysis of ecological momentary assessment (ema) data. Biometrics **64**(2), 627–634 (2008)
Hedeker, D., Demirtas, H., Mermelstein, R.J.: A mixed ordinal location scale model for analysis of ecological momentary assessment (ema) data. Stat. Interface **2**(4), 391 (2009)
Hedeker, D., Mermelstein, R.J., Demirtas, H.: Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. Stat. Med. **31**(27), 3328–3336 (2012)
Hornung, R.: Ordinal forests. J. Classif. **37**, 4–17 (2020)
Hothorn, T., Lausen, B.: On the exact distribution of maximally selected rank statistics. Comput. Stat. Data Anal. **43**, 121–137 (2003)
Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. **15**, 651–674 (2006)
Ishwaran, H., Gatsonis, C.A.: A general class of hierarchical ordinal regression models with applications to correlated roc analysis. Can. J. Stat. **28**(4), 731–750 (2000)
Janitza, S., Tutz, G., Boulesteix, A.L.: Random forests for ordinal responses: prediction and variable selection. Comput. Stat. Data Anal. **96**, 57–73 (2016)
Karlson, K.B., Holm, A., Breen, R.: Comparing regression coefficients between same-sample nested models using logit and probit: a new method. Sociol. Methodol. **42**(1), 286–313 (2012)
Loh, W.Y.: Fifty years of classification and regression trees. Int. Stat. Rev. **82**(3), 329–348 (2014)
Long, J.S., Allison, P.D., McGinnis, R.: Rank advancement in academic careers: sex differences and the effects of productivity. Am. Sociol. Rev. **58**(5), 703–722 (1993)
McCullagh, P.: Regression model for ordinal data (with discussion). J. R. Stat. Soc. B **42**, 109–127 (1980)
Mood, C.: Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur. Sociol. Rev. **26**(1), 67–82 (2010)
Murphy, A.H.: The ranked probability score and the probability score: a comparison. Weather **81**, 82 (1970)
Piccarreta, R.: Classification trees for ordinal variables. Comput. Stat. **23**(3), 407–427 (2008)
Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)
Rohwer, G.: A note on the heterogeneous choice model. Sociol. Methods Res. **44**(1), 145–148 (2015)
Shih, Y.S.: A note on split selection bias in classification trees. Comput. Stat. Data Anal. **45**, 457–466 (2004)
Shih, Y.S., Tsai, H.: Variable selection bias in regression trees with constant fits. Comput. Stat. Data Anal. **45**, 595–607 (2004)
Strobl, C., Boulesteix, A.L., Augustin, T.: Unbiased split selection for classification trees based on the gini index. Comput. Stat. Data Anal. **52**, 483–501 (2007)
Strobl, C., Malley, J., Tutz, G.: An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol. Methods **14**, 323–348 (2009)
Tutz, G.: Binary response models with underlying heterogeneity: identification and interpretation of effects. Eur. Sociol. Rev. **34**, 211–221 (2018)
Tutz, G., Berger, M.: Separating location and dispersion in ordinal regression models. Econom. Stat. **2**, 131–148 (2017)
Tutz, G., Berger, M.: Tree-structured modelling of categorical predictors in generalized additive regression. Adv. Data Anal. Classif. **12**, 737–758 (2018). https://doi.org/10.1007/s11634-017-0298-6
Williams, R.: Using heterogeneous choice models to compare logit and probit coefficients across groups. Sociol. Methods Res. **37**(4), 531–559 (2009)
Williams, R.: Fitting heterogeneous choice models with oglm. Stata J. **10**(4), 540 (2010)
Wright, M.N., König, I.R.: Splitting on categorical predictors in random forests. PeerJ **7**, e6339 (2019)
Zeileis, A., Hothorn, T., Hornik, K.: Model-based recursive partitioning. J. Comput. Graph. Stat. **17**(2), 492–514 (2008)