Journal of Classification
https://doi.org/10.1007/s00357-021-09406-4

Ordinal Trees and Random Forests: Score-Free Recursive Partitioning and Improved Ensembles

Gerhard Tutz[1]

Accepted: 16 November 2021Published online: 04 December 2021

Abstract
Existing ordinal trees and random forests typically use scores that are assigned to the ordered categories, which implies that a higher scale level is used. Versions of ordinal trees are proposed that take the scale level seriously and avoid the assignment of artificial scores. The construction principle is based on an investigation of the binary models that are implicitly used in parametric ordinal regression. These building blocks can be fitted by trees and combined in a similar way as in parametric models. The obtained trees use the ordinal scale level only. Since binary trees and random forests are constituent elements of the proposed trees, one can exploit the wide range of binary trees that have already been developed. A further topic is the potentially poor performance of random forests, which seems to have been neglected in the literature. Ensembles that include parametric models are proposed to obtain prediction methods that tend to perform well in a wide range of settings. The performance of the methods is evaluated empirically by using several data sets.

Keywords Recursive partitioning · Trees · Random forests · Ensemble methods · Ordinal regression

1 Introduction

There is a long tradition of analyzing ordinal response data by using parametric models, which started with the seminal paper of McCullagh (1980). Overviews on developments that include nonparametric approaches have been given, for example, by Agresti (2010) and Tutz (2020). More recently, recursive partitioning method have been developed that allow to investigate the impact of explanatory variables on ordinal responses by nonparametric tools. Single and random forests for ordinal responses have several advantages, they can be applied to large data sets and are considered to perform very well in prediction.

A problem with most of the ordinal trees is that they assume that scores are assigned to the ordered categories of the response. The assignment of scores can be warranted in some cases, in particular if ordinal responses are built from continuous variables by grouping.

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

[1] Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 München, Germany

However, it is rather artificial and arbitrary in genuine ordinal response data, for example, if the response represents ordered levels of severeness of a disease. Then, one can not choose the midpoints of the intervals from which the ordered response is built as suggested by Hothorn et al. (2006) since no continuous variable is observed. If nevertheless scores are assigned they can affect the prediction results although that has not always to be the case (see also Janitza et al., 2016).

The packages *rpartOrdinal* (Archer, 2010) as well as the improved version *rpartScore* (Galimberti et al., 2012), which are based on the Gini impurity function, use assigned scores. The same holds for the random forests proposed by Janitza et al. (2016) and the ordinal version of conditional trees of the package *party* (Hothorn et al., 2006; Hothorn & Zeileis, 2015). The random forest approach proposed by Hornung (2020) is somewhat different, it also translates ordinal measurements into continuous scores but optimizes scores instead of using a fixed score. Sciandra et al. (2017) investigated the correspondence between ranked preferences and data on an ordinal scale in the case of multivariate ordinal data. Versions of random forests without scores were proposed more recently by Buri and Hothorn (2020). They use the ordinal proportional odds model to obtain statistics that are used in splitting. The trees proposed by Cappelli et al. (2019) are based on the so-called CUB model, which has been reviewed by Piccolo and Simone (2019). An alternative semiparametric approach that uses parametric models has been proposed by Simone and Tutz (2020).

In the following alternative trees and random forests that take the scale level of the response seriously are proposed. The main concept is that ordinal responses contain binary responses as building blocks. This has already been implicitly used in parametric modeling approaches. For example, the widely used proportional odds model can be seen as a model that parameterizes the split of response categories into two groups of adjacent categories. But the principle also holds for alternative models as the adjacent categories model and the sequential model (see Tutz, 2020 for an overview and a taxonomy of ordinal regression models). The proposed trees explicitly use the representation of ordinal responses as a set of binary variables. Random forests for the binary variables are used to obtain random forests for ordinal response data.

For random forests it is important that they provide good performance in terms of prediction. They are commonly considered as being very efficient. However, as will be demonstrated this does not hold in general. In many cases simple parametric models turn out to be at least as efficient and sometimes more efficient than the carefully designed random forests. Typically, when ordinal forests are propagated the accuracy is investigated for versions of random forests only but they are not compared to parametric competitors. In the following we propose the use of ensembles that include parametric models to provide a stable prediction tool that works well in all kinds of data sets. For overviews on ensemble methods and multimodel inference (see, for example, Polikar, 2009; Burnham & Anderson, 2002).

The paper has two objectives, introducing score-free recursive partitioning and random forests, and proposing ensembles that include parametric models. In Section 2 the representation of ordinal responses as a sequence of binary responses is briefly considered. It makes clear that specific binary responses can be seen as building blocks of classical parametric models. In Section 3 it is shown how these building blocks can be used to construct score-free trees and random forests. In addition, more general ensembles are considered. In Section 4 the performance of the ensembles is investigated by using real data sets. Section 5 is devoted to importance measures, which are an essential ingredient of random forests since the impact of variables on prediction in random forests is not directly available.

2 Binary Representations of Ordinal Responses

In the following the representation of ordinal response as a collection of binary responses is considered. It can be seen as being behind the construction of parametric ordinal models and will serve to construct a novel type of recursive partitioning that does not use assigned scores.

Let the ordinal response Y take values from $\{1, \ldots, k\}$. Although these values suggest a univariate response the actual response is multivariate since the numbers $1, \ldots, k$ just represent that outcomes are ordered but distances between numbers assigned to categories should not be built in an ordinal scale because they are not interpretable.

A multivariate representation of the outcome can be obtained by using binary dummy variables. Natural candidates for dummy variables are the split variables

$$Y_r = \begin{cases} 1 & Y \geq r \\ 0 & Y < r, \end{cases} \tag{1}$$

$r = 1, , k$, where category 1 serves as a reference category, and $Y_1 \equiv 1$. Then, $Y = r$ is represented by a sequence of $r - 1$ ones followed by a sequence of zeros,

$$(Y_2 \ldots, Y_k) = (1, \ldots, 1, 0, \ldots, 0).$$

The vector $(Y_2 \ldots, Y_k)$ can be seen as a multivariate representation of the response. The dummy variables that generate vectors of this form, which are characterized by a sequence of ones followed by a sequence of zeros have also be referred to as Guttman variables (Andrich, 2013).

Classical ordinal regression model use these dummy variables but are most often derived from the assumption of an underlying continuous variable, and the link to split variables is ignored. The most widely used *proportional odds model*, also called *cumulative logistic model*, has the form

$$P(Y \geq r|\boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}), \quad r = 2, \ldots, k. \tag{2}$$

where $\boldsymbol{x}$ is a vector of explanatory variables and $F(\eta) = \exp(\eta)/(1+\exp(\eta))$ is the logistic distribution function. For the parameters one has the restriction $\beta_{02} \geq \cdots \geq \beta_{0k}$. The model explicitly uses the dichotomizations given by (1). Since $Y \geq r$ iff $Y_r = 1$ the model can also be given as

$$P(Y_r = 1|\boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}), \quad r = 2, \ldots, k. \tag{3}$$

Thus, the proportional odds model is equivalent to a collection of binary logit models that have to hold simultaneously. The model implies that the effect of covariates contained in $\boldsymbol{x}^T\boldsymbol{\beta}$ is the same for all dichotomizations. That means if one fits the binary models (3) separately one should obtain similar values for estimates of $\boldsymbol{\beta}$. This restriction can be weakened by using the partial proportional odds model, in which the effect of variables may depend on the category, that is, the linear term $\boldsymbol{x}^T\boldsymbol{\beta}$ in (2) is replaced by $\boldsymbol{x}^T\boldsymbol{\beta}_r$. However, as will be discussed later the parameters $\boldsymbol{\beta}_r$ can not vary freely.

Model (2) is a so-called cumulative model since on the left hand side one has the sum of probabilities $P(Y \geq r|\boldsymbol{x})$. Cumulative models form a whole family of models, whose members are characterized by the choice of a specific strictly increasing distribution function $F(.)$. They have been investigated and extended, among others, by McCullagh (1980), Brant (1990), Peterson and Harrell (1990), Bender and Grouven (1998), Cox (1995), and Kim (2003) and Liu et al. (2009).

An alternative ordinal regression model is the *adjacent categories model*, which has the basic form

$$P(Y \geq r|Y \in \{r-1, r\}, \boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}), \quad r = 2, \ldots, k. \tag{4}$$

Since $P(Y \geq r|Y \in \{r-1, r\}, \boldsymbol{x}) = P(Y = r|Y \in \{r-1, r\}, \boldsymbol{x})$ it specifies the probability of observing category r given the response is in categories $\{r-1, r\}$. Because of the conditioning it can be seen as a local model. The interesting point is that it also uses the split variables. It is easily seen that it is equivalent to

$$\begin{aligned} &P(Y_r = 1|Y_{r-1} = 1, Y_{r+1} = 0, \boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}), \quad r = 2, \ldots, k-1, \\ &P(Y_k = 1|Y_{k-1} = 1, \boldsymbol{x}) = F(\beta_{0k} + \boldsymbol{x}^T\boldsymbol{\beta}). \end{aligned} \tag{5}$$

Thus, it specifies the binary response variable Y_r *conditionally* in contrast to cumulative models, which determine the binary response directly in an unconditional way. But as for cumulative models it is assumed that the binary models (5) hold simultaneously.

The adjacent categories logit model may also be considered as the regression model that is obtained from the row-column (RC) association model considered by Goodman (1981a), Goodman (1981b), and Kateri (2014). It is also related to Anderson's stereotype model (Anderson, 1984), which was considered by Greenland (1994) and Fernandez et al. (2019). It has been most widely used as a latent trait model in the form of the partial credit model (Masters, 1982; Masters & Wright, 1984; Muraki, 1997).

An advantage of the adjacent categories model is that one can replace the parameter vector $\boldsymbol{\beta}$ by a category-specific parameter vector $\boldsymbol{\beta}_r$ without running into problems. In cumulative models one has the restriction $P(Y \geq 2|\boldsymbol{x}) \geq \cdots \geq P(Y \geq k|\boldsymbol{x})$, which can yield problems, in particular when fitting the binary models (3) with category-specific parameter vectors $\boldsymbol{\beta}_r$. For overviews of parametric ordinal models (see, for example, Agresti, 2010; Tutz, 2012). They also include a third type of ordinal model, the sequential model, which is a specific process model, which could also be extended to tree type models. But because of its specific nature we do not consider it explicitly.

The main point is that binary models are at the core of parametric classical ordinal models. There is a good reason for that because the splits represent the order in categories without assuming more than an order of categories. In the next section this is exploited to construct trees that account for the ordering of categories. It should, nevertheless, be noted that alternative models for ordinal responses have been proposed (see, for example, Biernacki & Jacques, 2016; Ursino & Gasparini, 2018; Piccolo & Simone, 2019).

It should be noted that the approach of dichotomizing outcomes has been used before for continuous outcomes, an early reference is Foresi and Peracchi (1995). It allows flexible models for conditional distribution functions to be fitted by application of relatively simple models for binary outcomes, and can be used in regression models that aim at estimating the whole conditional distribution (see, for example, Chernozhukov et al., 2013). The approaches typically use unconditional binary models. The strength of the modeling approach used here is that conditional binary models avoid the need for an explicit monotonicity constraint.

3 Recursive Partitioning Based on Splits

The crucial role of split variables in modeling ordered response can be used to obtain nonparametric tree models that use the ordering efficiently. There are basically two ways to do so, one is by using the split variables directly, which corresponds to cumulative type

models, the other approach is to use them conditionally, which corresponds to the adjacent categories approach.

3.1 Trees for Split Variables

Split variables are binary, and therefore, binary trees can be fitted. Let the tree for Y_r be given by

$$\log \frac{P(Y_r = 1|\boldsymbol{x})}{P(Y_r = 0|\boldsymbol{x})} = \text{tr}_r(\boldsymbol{x}), \quad r = 2, \ldots, k, \tag{6}$$

where $\text{tr}_r(\boldsymbol{x})$ denotes the partitioning of the predictor space, that is, the tree. Then, one obtains for the probabilities

$$P(Y_r = r|\boldsymbol{x}) = P(Y \geq r|\boldsymbol{x}) = \frac{\exp(\text{tr}_r(\boldsymbol{x}))}{1 + \exp(\text{tr}_r(\boldsymbol{x}))}.$$

The corresponding trees are called *split-based trees*. Split variables are a formal tool to group categories but have substantial meaning in many applications. For example, in the retinopathy data set (Bender & Grouven, 1998), which will also be considered later, the response categories are (1) no retinopathy, (2) nonproliferative retinopathy, and (3) advanced retinopathy or blind. Thus, the split between categories $\{1\}$ and $\{2, 3\}$ distinguishes between healthy and not healthy, whereas the split between $\{1, 2\}$ and $\{3\}$ distinguishes between serious illness and otherwise. It is crucial that explanatory variables may play different roles for different splits. In the retinopathy data set, with explanatory variables smoking (SM = 1: smoker, SM = 0: non-smoker), diabetes duration (DIAB) measured in years, glycosylated hemoglobin (GH), measured in percent, and diastolic blood pressure (BP) measured in mmHg, one obtains for the two splits the trees shown in Fig. 1 (fitted by using *ctree*, Hothorn et al. (2006)). It is seen that trees are quite different, which means that explanatory variables play differing roles when used to distinguish between healthy and not healthy and between serious illness and less serious illness.

3.2 Trees for Conditional Splits

Instead of the unconditional split variables considered previously let us consider the conditional binary variables

$$\tilde{Y}_r = \begin{cases} 1 & Y \geq r \text{ given } Y \in \{r-1, r\} \\ 0 & Y < r \text{ given } Y \in \{r-1, r\}, \end{cases} \tag{7}$$

$r = 2, \ldots, k$. The variables are conditional versions of split variables. More concrete, $\tilde{Y}_r$ represents $Y_r|Y_{r-1} = 1, Y_{r+1} = 0$. The main difference between $\tilde{Y}_r$ and Y_r is that the former is a conditional variable. This is important since fitting a tree to $\tilde{Y}_r$ means one includes only observations with $Y \in \{r-1, r\}$. The corresponding tree can be seen as a nonparametric version of the adjacent categories model and is called an *adjacent categories tree*. The corresponding trees are local, they reflect the impact of explanatory variables on the distinction between adjacent categories.

Adjacent categories trees have a different interpretation than trees for split variables. For illustration, Fig. 2 shows the fitted trees for the retinopathy data. It is seen that diabetes duration (DIAB) has an impact in both trees. In the split between categories 1 and 2 the only other variable that is significant is glycosylated hemoglobin while in the split between categories 2 and 3 it is blood pressure. Trees are smaller than split-based trees since due

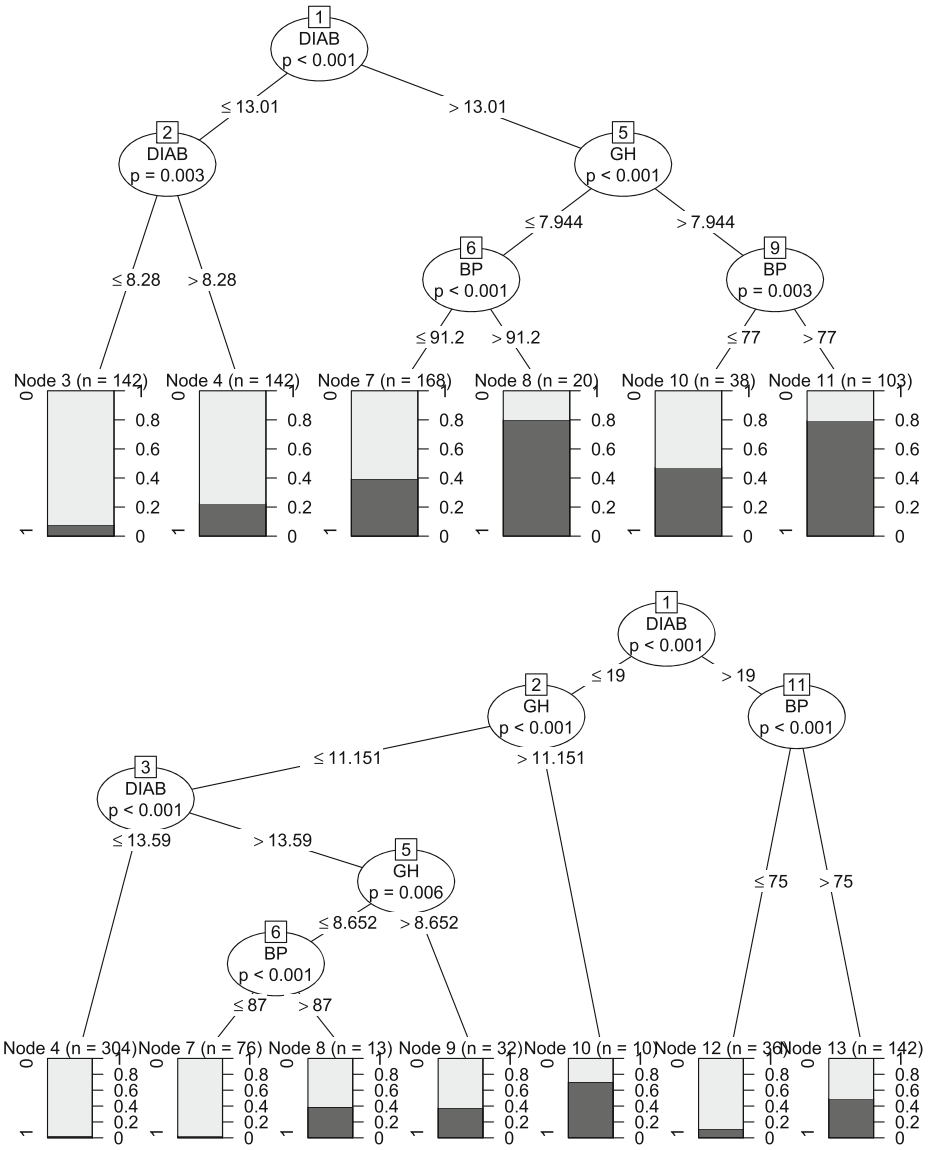

Fig. 1 Conditional trees for retinopathy data, upper panel: split between {1} and {2, 3}. lower panel: split between {1, 2} and {3}

to conditioning the number of observations is smaller. From a substantial point of view it might be most interesting to combine trees from the different splitting concepts. The first tree in Fig. 1 distinguishes between {1} and {2, 3}, that is between healthy and non healthy. The second tree in Fig. 2 shows which variables are significant when distinguishing between categories 2 and 3 given the response is in categories {2, 3}, that is, which variables are influential given the patient suffers from retinopathy.

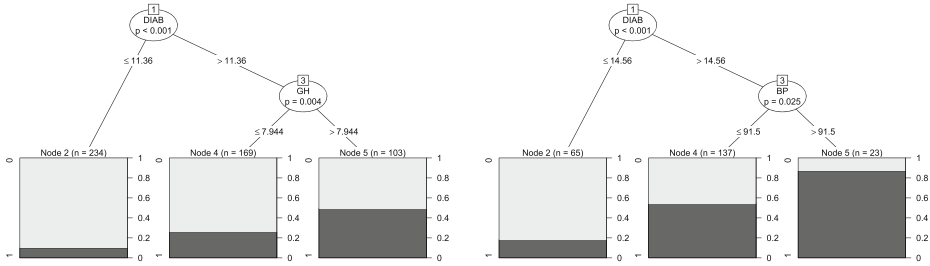

Fig. 2 Conditional trees for retinopathy data, left: split given $Y \in \{1, 2\}$, right: split given $Y \in \{2, 3\}$

3.3 From Trees to Random Forests

Single trees can be informative for researchers that want to investigate which variables have an impact on specific dichotomizations. If one has prediction in mind a better choice are random trees, which are much more stable and efficient than single trees (Breiman, 1996; 2001; Bühlmann et al. 2002). Then, it is necessary to combine the results of single trees in a proper way.

Let us first consider split-based trees. They face the problem familiar from cumulative models with category-specific effects that specific constraints have to be fulfilled. More specifically, for all values of $\boldsymbol{x}$ the constraint $P(Y \geq 2|\boldsymbol{x}) \geq \cdots \geq P(Y \geq k-1|\boldsymbol{x})$ has to hold, which is equivalent to $P(Y_2 = 1|\boldsymbol{x}) \geq \cdots \geq P(Y_{k-1} = 1|\boldsymbol{x})$. However, for separately fitted trees the corresponding condition $\text{tr}_2(\boldsymbol{x}) \geq \cdots \geq \text{tr}_{k-1}(\boldsymbol{x})$ does no necessarily hold. The same problem occurs in partial proportional odds model, for which $\beta_{02} + \boldsymbol{x}^T\boldsymbol{\beta}_2 \geq \cdots \geq \beta_{0k} + \boldsymbol{x}^T\boldsymbol{\beta}_k$ has to hold.

Let $\hat{\pi}(\boldsymbol{x})^{(r)} = \hat{P}(Y \geq r|\boldsymbol{x})$ denote the estimated cumulative probabilities resulting from the tree for the split variable Y_r. Then, probabilities are obtained by $\hat{P}(Y = r|\boldsymbol{x}) = \hat{\pi}(\boldsymbol{x})^{(r)} - \hat{\pi}(\boldsymbol{x})^{(r+1)}$ if $\hat{\pi}(\boldsymbol{x})^{(r)} \geq \hat{\pi}(\boldsymbol{x})^{(r+1)}$ for all r. If the latter condition does not hold cumulative probabilities $\hat{\pi}(\boldsymbol{x})^{(r)}, \ldots, \hat{\pi}(\boldsymbol{x})^{(k)}$ are fitted to be decreasing by using monotone regression tools. Alternative approaches to obtain compatible estimators have been considered in the machine learning community, for example, by Chu and Keerthi (2007).

An advantage of adjacent categories trees is that no monotonization tools are needed since estimated probabilities are always compatible. Let the adjacent categories trees be given by

$$\log \frac{P(\tilde{Y}_r = 1|\boldsymbol{x})}{P(\tilde{Y}_r = 0|\boldsymbol{x})} = \tilde{\text{tr}}_r(\boldsymbol{x}), \quad r = 2, \ldots, k, \tag{8}$$

where $\tilde{\text{tr}}_r(\boldsymbol{x})$ denotes the partitioning of the predictor space. It is not hard to derive that the probability of an response in category r given the representation (8) holds has the form

$$P(Y_r = r|\boldsymbol{x}) = \frac{\exp(\sum_{s=2}^{r} \tilde{\text{tr}}_s(\boldsymbol{x}))}{\sum_{s=1}^{k} \exp(\sum_{l=2}^{s} \tilde{\text{tr}}_s(\boldsymbol{x}))}, \tag{9}$$

where $\sum_{l=2}^{1} \tilde{\text{tr}}_s(\boldsymbol{x})) = 0$. The representation (9) holds for any values of $\tilde{\text{tr}}_2(\boldsymbol{x}), \ldots, \tilde{\text{tr}}_k(\boldsymbol{x})$, no specific restriction has to be fulfilled.

Random forests are obtained by combining not only the trees for split variables but averaging over a multitude of trees generated by randomization. More concrete, for the split variables binary random forests are fitted and the prediction is combined by using (8) and (9), respectively. The approach exploits the role of the split variables as building

blocks for ordinal responses, and can be seen as a *split variables based* approach, which is unconditional in split-based trees and conditional in adjacent categories trees.

3.4 Ensemble Learners Including Parametric Models

Before investigating the proposed random forests in detail let us point out a problem with ordinal trees that is often ignored. Most presentations of ordinal trees focus on the development of novel trees but do not compare the performance of random forests to the performance of simple parametric models as the proportional odds model. That leaves the impression that random forests are the most efficient tools. As will be demonstrated in the following sections parametric models should not be ignored, in many applications they can perform as well as random forests or even better. The use of parametric ordinal models for the prediction of ordinal responses has some tradition (see, for example, Rudolfer et al., 1995; Campbell & Donner, 1989; Campbell et al., 1991; Anderson & Phillips, 1981).

Trees themselves are ensemble methods that combine various splits to obtain a good approximation of the underlying response probabilities. To exploit the potential strength of parametric models we propose an ensemble that includes these models. When estimating response probabilities we will use the ensemble

$$\hat{P}(Y = r|\boldsymbol{x}) = \sum_{j=1}^{M} w_j \hat{P}_j(Y = r|\boldsymbol{x}), \tag{10}$$

where $\hat{P}_j(Y = r|\boldsymbol{x})$ are estimated probabilities for the jth learner. Learners can be random forests but also parametric models. The weights w_j are chosen according to the prediction performance of the jth learner. More concrete, let $s_1, \ldots, s_M$ denote error scores for the M models or estimation methods. With $\min = \min\{s_1, \ldots, s_M\}$, $\max = \max\{s_1, \ldots, s_M\}$ the un-standardized weight for method i is defined by

$$\tilde{w}_i = a \times (s_i - \max) + \min,$$

where $a = \min \times (1 - M)/(\max - \min)$, such that the method with the largest error obtains weight min, and the method with the smallest value obtains weight $M \times \min$. The final weight is the standardized version $w_i = \tilde{w}_i / \sum_j \tilde{w}_j$. As error score we used the quadratic or Brier score, which is explicitly given in Section 4.1. For concrete data sets the weights are determined by splitting the data set several times into a learning data set that contains 70% of the data and using the rest of the data as validation sample in which the error scores and the weights are computed. Averaging of weights yields the final weights. We used only three splits, since higher numbers of splits did not improve the performance.

The ensemble efficiently uses different types of learners. By combining them it yields more stable predictions than single learners and automatically gives more weight to the best learner in the ensemble. One might use different error scores and different weighting schemes, but the weighting scheme used here which lets the un-standardized weights vary between min and $M \times \min$ showed rather good performance.

Typically, in classification predictions of single trees from an ensemble are combined by voting. Each subject with given values of the predictor is dropped through every tree such that each single tree returns a predicted class. The prediction of the ensemble is the class most trees voted for. One obtains a majority vote, which has also been called a committee method. It should be noted that the ensembles proposed here combine *probabilities*. They are not ensembles that use majority votes to combine class predictions obtained for each single learner. We also considered majority votes that combine the votes on splits but the results

were distinctly inferior to using probabilities. By computing the predicted class probabilities one can use more general accuracy measures that also take into account the precision of the prediction and obtain a better approximation to the true conditional distribution of responses.

While the use of parametric models in ensembles seems to have been neglected, there are several proposals how to form ensembles from tress (see, for example, the weighted random forests proposed by Winham et al. (2013) and the ensembles considered by Khan et al. (2020)). Also the generalized random forests proposed by Athey et al. (2019) combine alternative estimators, in their case forest-based estimates are obtained by local estimating equations, which uses that random forests can be viewed as locally weighted estimators as suggested by Hothorn et al. (2004) and also used by Meinshausen (2006).

4 Ordinal Random Forests and Prediction

4.1 Measuring Accuracy of Prediction

One way to investigate the power of a model is to investigate its ability to predict future observations. In discriminant analysis one often uses class prediction as a measure of performance. Class prediction in the considered framework comes in two forms. As predicted class one may use the mode of the response, $\hat{Y} = \text{mod}(\boldsymbol{x})$, which is in accordance with the Bayes prediction rule, or the median $\hat{Y} = \text{med}(\boldsymbol{x})$, which makes use of the ordering of categories. Then, for a new observation $(Y_0, \boldsymbol{x}_0)$, one typically considers the 0-1 loss function

$$L_{01}(Y_0, \hat{Y}_0) = I(Y_0 \neq \hat{Y}_0),$$

where $I(.)$ is the indicator function. One obtains 1 if the prediction is wrong, and 0 if the prediction is correct. The average over new observations yields the 0-1 error rate.

Rather than giving just one value as a predictor for the class it is more appropriate to consider the whole vector $\hat{\boldsymbol{pi}}^T(\boldsymbol{x}) = (\hat{\pi}_1(\boldsymbol{x}), \ldots, \hat{\pi}_k(\boldsymbol{x}))$, where $\hat{\pi}_r(\boldsymbol{x}) = P(Y_r = r|\boldsymbol{x})$ is the probability one obtains after fitting a tree. The vector $\hat{\boldsymbol{pi}}(\boldsymbol{x})$ represents the predictive distribution. As Gneiting and Raftery (2007) postulated a desirable predictive distribution should be as sharp as possible and well calibrated. Sharpness refers to the concentration of the distribution and calibration to the agreement between distribution and observation.

Since the response is measured on an ordinal scale an appropriate loss function derived from the *continuous ranked probability score* (Gneiting and Raftery (2007)) is

$$L_{RPS}(Y_0, \hat{\boldsymbol{pi}}) = \sum_{r=1}^{k} (\hat{\pi}(r, \boldsymbol{x}_0) - I(Y_0 \leq r))^2,$$

where $(Y_0, \boldsymbol{x}_0)$ is a new observation and $\hat{\pi}(r, \boldsymbol{x}_0) = \hat{\pi}_1(\boldsymbol{x}_0) + \cdots + \hat{\pi}_r(\boldsymbol{x}_0)$ is the cumulative probability. It takes the closeness between the whole distribution and the observed value into account (see Gneiting and Raftery (2007) for a discussion of its properties).

Further measures that use more information than the simple misclassification rate are the quadratic and the logarithmic score. The *quadratic score*, which is also known as *Brier score*, is given by

$$L_B(Y_0, \hat{\boldsymbol{pi}}) = (1 - \hat{\pi}_{Y_0}(\boldsymbol{x}_0))^2 + \sum_{r \neq Y_0} \pi_r(\boldsymbol{x}_0)^2.$$

It also measures the discrepancy between the true response and the estimated probabilities taking into account all of the k estimated probabilities. It is the empirical version of the quadratic loss function $L_2(\boldsymbol{pi}, \hat{\boldsymbol{pi}}) = \sum_r (\pi_r - \pi_r)^2$, where $\boldsymbol{pi}^T = (\pi_1, \ldots, \pi_k)$ is the vector of true probabilities and $\hat{\boldsymbol{pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_k)$ is the vector of estimated probabilities. The quadratic score results when using the degenerate vector $\boldsymbol{pi}^T = (0, \ldots, 0, 1, 0, \ldots, 0)$, which contains a single 1 in the observed category. For simplicity, in $L_B(Y_0, \hat{\boldsymbol{pi}})$ the observation that generates the degenerate vector of probabilities is used as argument. There is a strong link to the ranked probability score $L_{RPS}(Y_0, \hat{\boldsymbol{pi}})$, which can be seen as a sum of quadratic scores for the split variables.

A measure, which has also been used to evaluate the performance of predictors, is the logarithmic score given by

$$L_{log}(Y_0, \hat{\boldsymbol{pi}}) = -\log(\hat{\pi}_{Y_0}(\boldsymbol{x}_0)).$$

It is the empirical version of the Kullback-Leibler distance $L_{KL}(\boldsymbol{pi}, \hat{\boldsymbol{pi}}) = \sum_r \pi_r log(\pi_r / \hat{\pi}_r)$. While the Kullback-Leibler distance uses the whole distribution, the empirical version uses only the estimated probability of the observed category.

In the evaluations we use the measures that use the estimated probabilities since they contain more information than the simple 0-1 loss that yields the misclassification rate. The discrepancy between observations in the validation sample and estimated probabilities, which are obtained from the learning sample, always uses the estimated vector of probabilities. That means, in particular, that in the ensemble methods the probabilities (10) are used.

4.2 Data Sets

4.2.1 Heart Data

This data set includes 294 patients undergoing angiography at the Hungarian Institute of Cardiology in Budapest between 1983 and 1987, and is included in the R package *ordinalForest* (Hornung, 2020). It contains ten covariates and one ordinal target variable. Explanatory variables are age (age in years), sex (1 = male; 0 = female), chest pain (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic), trestbps (blood pressure in mm Hg on admission to the hospital), chol (serum cholestoral in mg/dl), fbs (fasting blood sugar > 120 mg/dl, 1 = true; 0 = false) restecg (resting electrocardiographic results, 1 = having ST-T wave abnormality, 0 = normal), thalach (maximum heart rate achieved), exang (exercise induced angina, 1 = yes; 0 = no), oldpeak (ST depression induced by exercise relative to rest). The response is Cat (severity of coronary artery disease determined using angiograms, 1 = no disease; 2 = degree 1; 3 = degree 2; 4 = degree 3; 5 = degree 4).

4.2.2 Wine Data

We use the wine quality data (winequality-white) available from the UCI Machine Learning Repository (see also Cortez et al., 2009). The response is the quality in wine in ordered categories. Explanatory variables are fixed acidity, volatile acidity, citric acid, residual sugar,

chlorides and free sulfur dioxide. The original scoring of the quality is between 0 and 10, but not all of the categories have been used leaving 5 response categories.

4.2.3 Housing Data

We use the housing data for 506 census tracts of Boston in the version BostonHousing2, which contains the corrected version of the original data by Harrison and Rubinfeld (1978) and included additional spatial information. It is included in the R package *mlbench*. As categorical response we use the corrected median value of owner-occupied homes in USD 1000's (cmedv) by binning the variable according to the cutoffs: 15, 19, 22, 25, and 32. Explanatory variables are crim (per capita crime rate by town, var 1), lstat (percentage of lower status of the population, var 2) zn (proportion of residential land zoned for lots over 25,000 sq. ft, var 3), nox (nitric oxides concentration inparts per 10 million, var 4), rm (average number of rooms per dwelling, var 5), dis (weighted distances to five Boston employment centres, var 6), rad (index of accessibility to radial highways, var 7), tax (full-value property-tax rate per USD 10,000, var 8), ptratio (pupil-teacher ratio by town, var 9), b (proportion of blacks by town, var 10), indus (proportion of non-retail business acres per town, var 11), age (proportion of owner-occupied units built prior to 1940, var 12).

4.2.4 Birth Weight Data

The *lobwt* data set contained in the R package *rpartOrdinal* has been used in several random forest papers. As categorical response we use the birth weight by binning the variable bwt according to the cutoffs: 2500,3000, and 3500 (see also Galimberti et al., 2012). Explanatory variables are age (age of mother in years), lwt (weight of mother at last menstrual period in Pounds), smoke (Smoking status during pregnancy, 1: No, 2: Yes), ht (history of hypertension, 1: No, 2: Yes), ftv (number of physician visits during the first trimester,1: None, 2: One, 3: Two, etc)

4.2.5 Retinopathy Data

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

4.2.6 Medical Care

Deb and Trivedi (1997) analyzed the demand for medical care for individuals, aged 66 and over, based on a data set from the U.S. National Medical Expenditure survey in 1987/88. The data ("NMES1988") are available from the R package AER (Kleiber & Zeileis, 2008). We consider the number of physician/non-physician office and hospital outpatient visits as outcome variable binning the variable according to the cutoffs: 0, 1, 3, 6, 8 and 11. The covariates used in the present analysis are the number of emergency room visits (emergency), the number of hospital stays (Hosp), the self-perceived health status (Health; 0: poor, 1: excellent), the number of chronic conditions (Numchron), a factor (adl) indicating

whether the individual has a condition that limits activities of daily living ("limited") or not ("normal"). age, maritial status (Married; 0: no, 1: yes), is the individual African-American (afam), emlpoyment satus(employes), and is the individual covered by private insurance (insurance). Since the effects vary across gender, we consider male patients only. F

4.2.7 GLES Data

The GLES data stem from the German Longitudinal Election Study (GLES), which is a long-term study of the German electoral process (Rattinger et al., 2014). The data consist of 2036 observations and originate from the pre-election survey for the German federal election in 2017 and are concerned with political fears. In particular the participants were asked: "How afraid are you due to the use of nuclear energy? The answers were measured on Likert scales from 1 (not afraid at all) to 7 (very afraid). The explanatory variables in the model are *Abitur* (high school certificate, 1: Abitur/A levels; 0: else), *Age* (age of the participant), *EastWest* (1: East Germany/former GDR; 0: West Germany/former FRG), *Gender* (1: female; 0: male), *Unemployment* (1: currently unemployed; 0: else).

4.2.8 Safety Data

The package CUB (Iannario et al., 2020) contains the data set relgoods, which provides results of a survey aimed at measuring the subjective extent of feeling safe in the streets. The data were collected in the metropolitan area of Naples, Italy. Every participant was asked to assess on a 10 point ordinal scale his/her personal score for feeling safe with large categories referring to feeling safe. There are $n = 2225$ observations and five variables, *Age*, *Gender* (0: male, 1: female), the educational degree (*EduDegree*; 1: compulsory school, 2: high school diploma, 3: Graduated-Bachelor degree, 4: Graduated-Master degree, 5: Post graduated), *WalkAlone* (1 = usually walking alone, 0 = usually walking in company), *Residence* (1: City of Naples, 2: District of Naples, 3: Others Campania, 4: Others Italia).

4.3 Ensembles at Work

In the following the accuracy of prediction in the data sets described above is investigated. The data sets were split repeatedly into a learning set with sample size n_L and a validation set built from the rest of the data (number of splits: 30). The learning set was used to fit the method under investigation, the accuracy of prediction is then computed in the validation set. We use all the accuracy measures that contain more information than simple class predictions. In addition, we give, for simplicity, the Euclidean distance between the predicted class and the true class, where the predicted class is determined by the median of the estimated probabilities. The latter measure is an indicator how far the prediction is from the true class.

The fitting of split-based and adjacent categories random forests can be based on different random forest methods for binary responses. In particular one can use *ordinalForest* (Hornung, 2020), *randomforest* (Liaw et al., 2015), or conditional trees as provided by *cforest* (Hothorn & Zeileis, 2015). Figure 3 shows the averaged ranked probability scores for fitted adjacent categories random forests when using ordinalForest (OrdRF), randomforest (RF), and cforest (CRF) for the housing data and the GLES data. The differences in performance are negligible. Therefore, in the following we use only one method to generate split-based and adjacent categories random forests, namely *randomforest*, which is computationally quite efficient.

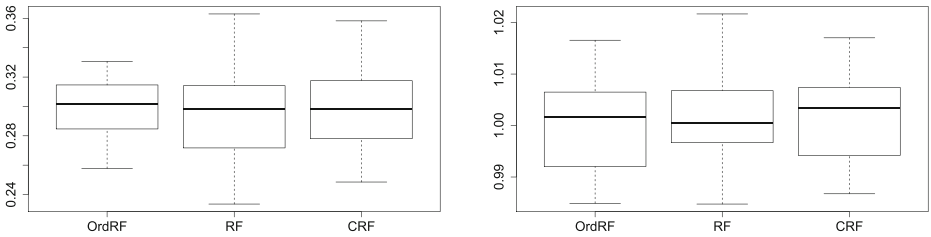

Fig. 3 Ranked probability scores for housing data (left) and GLES data (right) when fitting the adjacent categories random forest with ordinalForest (OrdRF), randomforest (RF), and cforest (CRF)

The methods to be considered in the following are:

- *Pom*: fitting of a proportional odds model,
- *Adj*: fitting of an adjacent categories logit model,
- *RFord*: fitting of an ordinal forest with *ordinalForest* as proposed by Hornung (2020),
- *RFT*: Ordinal random forest based on transformation models using *traforest* from package *trtf*, proposed by Buri and Hothorn (2020),
- *RFSp*: split-based ordinal random forest using *randomForest* to fit the binary random forests,
- *RFadj*: fitting of an adjacent categories random forest using *randomForest* to fit the binary random forests,
- *Ens3*: weighted ensemble including the proportional odds model, ordinalForest fit and adjacent categories random forest,
- *Ens5*: weighted ensemble including the proportional odds model, the adjacent categories model, ordinalForest fit, and adjacent categories and split-based random forest.

The first two methods use parametric models. The methods *RFord* and *RFT* are ordinal trees that have been proposed more recently. The next two methods, *RFSp* and *RFSp* are split based, which combine random forests for split variables. The last two methods are ensemble methods that include parametric models. *Ens3* is built from one parametric model, an ordinal random forest, and the adjacent categories random forest, whereas *Ens5* contains in addition the adjacent categories model and the split-based random forest. The ensemble built from three methods serves to demonstrate that it is essential to combine ordinal random forests and parametric models. The inclusion of further models will be shown to improve the performance only slightly. Since the number of tuning parameters vary across methods optimization of tuning parameters might favor the ones that allow for more tuning parameters. Therefore, we use for the methods the default values.

Figures 4, 5, 6, 7, 8, 9, 10 and 11 show the accuracy measures obtained for the validation data. For the unconditional split approach *RFSp* log scores are typically not available since estimated probabilities for some of the categories are close to zero, and therefore are not shown. It is seen from the plots that ordinal random forests outperform parametric models for the first two data sets. In both data sets random forests perform distinctly better. Among the methods proposed in the literature more recently *RFord* shows better performance than *RFT*. For the first data set the new split-based procedures performs as well as *RFord*, in the second the performance is closer to *RFT*.

In the next four data sets the performance of all the methods is comparable. In particular the parametric models do not perform worse or better than random forests approaches. In

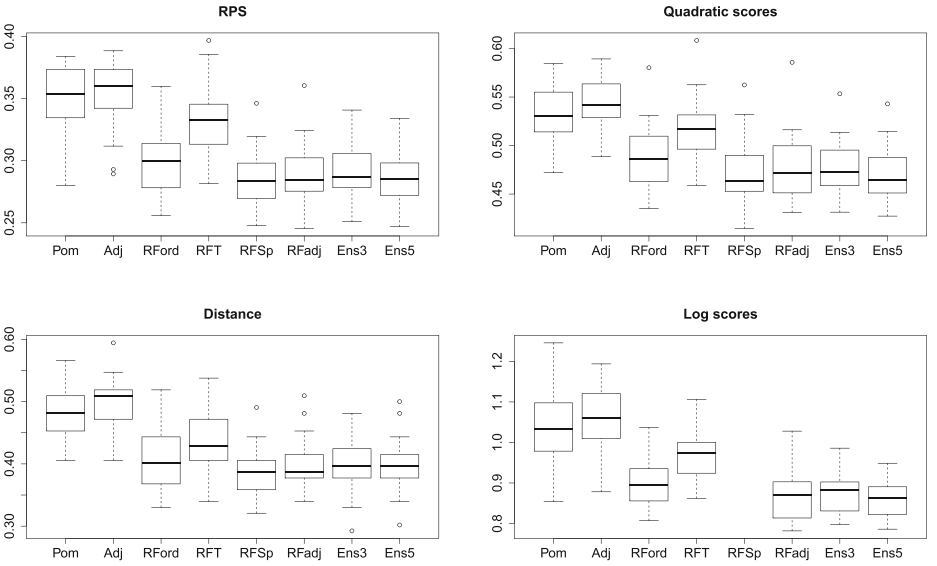

Fig. 4 Results for housing data ($n_L = 400$, six response categories)

two of the data sets, retinopathy and medical care, the methods *RFord* and *RFT* tend to perform slightly better than the split-based approaches.

In the third group of data sets, the GLES and the safety data, the parametric models show much better performance than the random forests methods. The random forests method that comes closest to the performance of the parametric model is *RFT*. There is a good reason for that since *RFT* implicitly uses a cumulative model. Therefore, in cases in which the

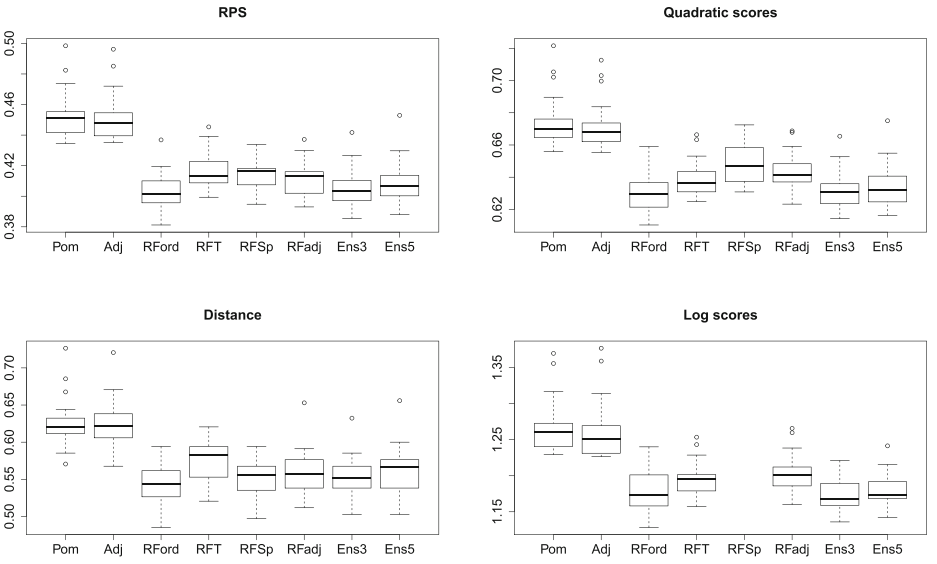

Fig. 5 Results for wine data ($n_L = 160$, five response categories)

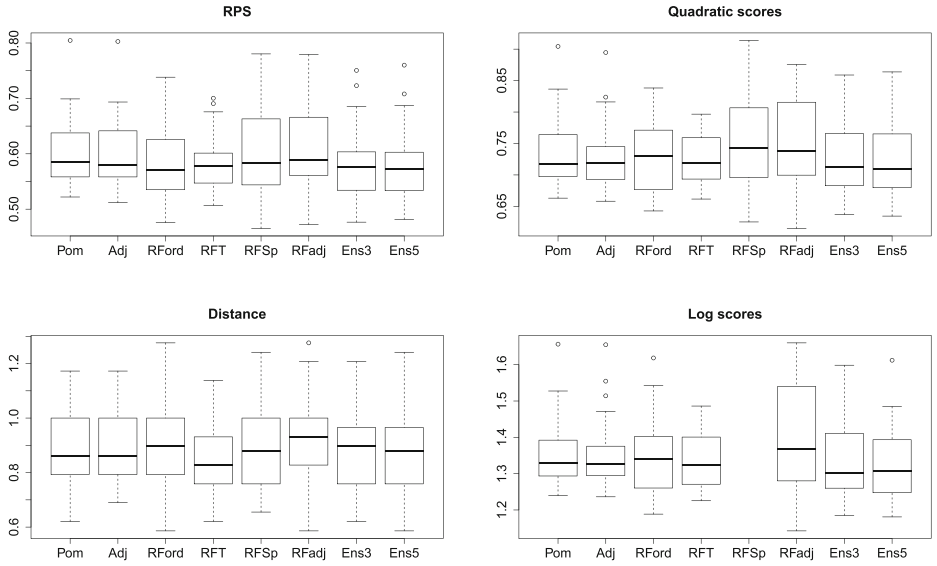

Fig. 6 Results for birth data ($n_L = 160$, four response categories)

cumulative model performs better than most random forests the *RFT* method should also do well. In particular split-based approaches are not the best choice if parametric models show good performance.

As far as parametric models and random forests are concerned, the performance depends on the data set. There are data sets in which the random forests have distinct advantages over parametric models. Then, split-based approaches show good performance. In surprisingly

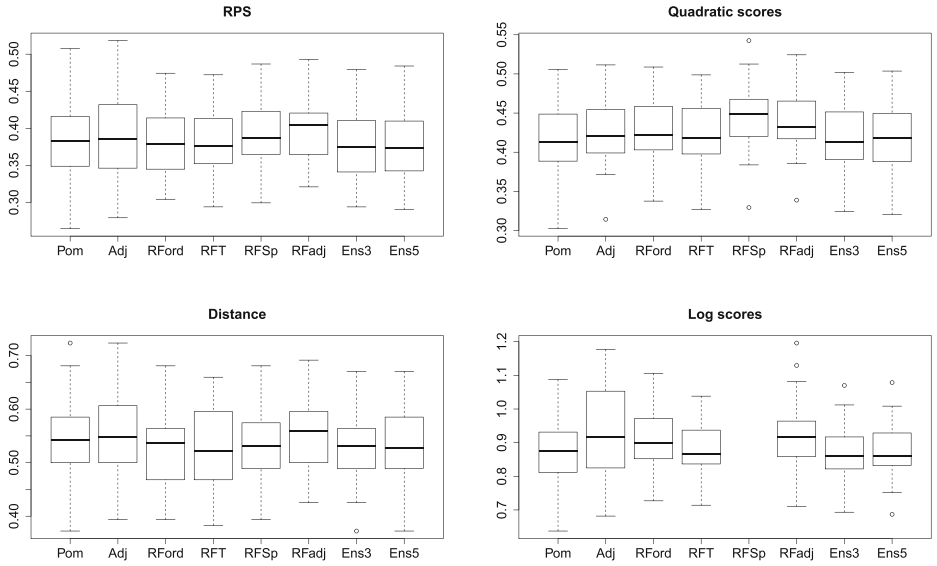

Fig. 7 Results for heart data ($n_L = 200$, five response categories)

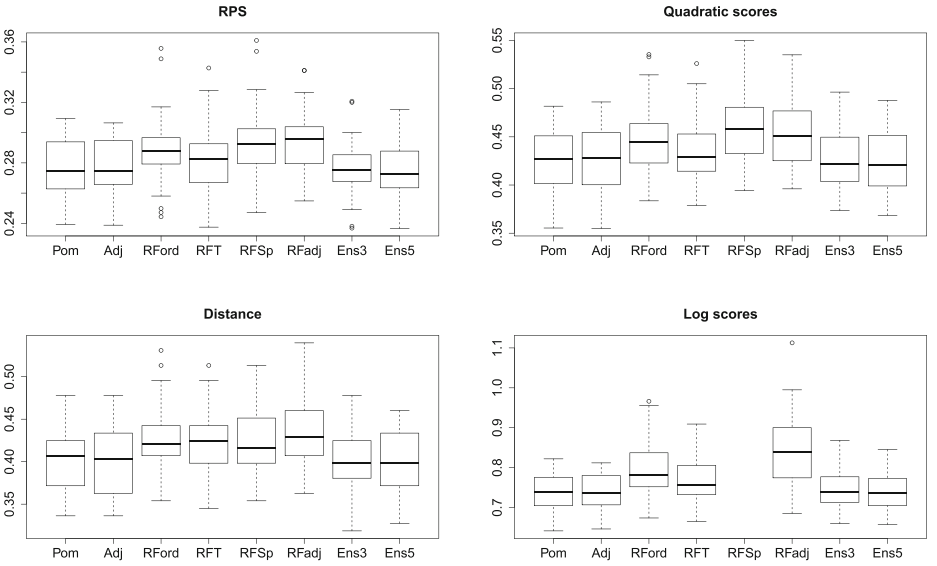

Fig. 8 Results for retinopathy data ($n_L = 500$, three response categories)

many data sets considered here random forests had no advantage over parametric models and there is not too much difference in the performance of methods. In a third type of data parametric models are distinctly to be preferred, both data sets for which this was found were questionnaire data.

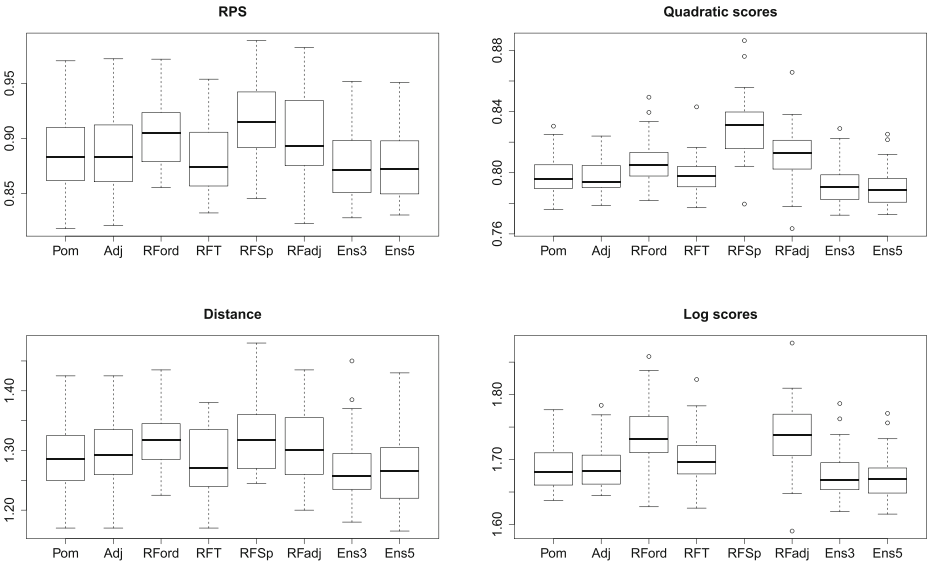

Fig. 9 Results for medical care data ($n_L = 300$, seven response categories)

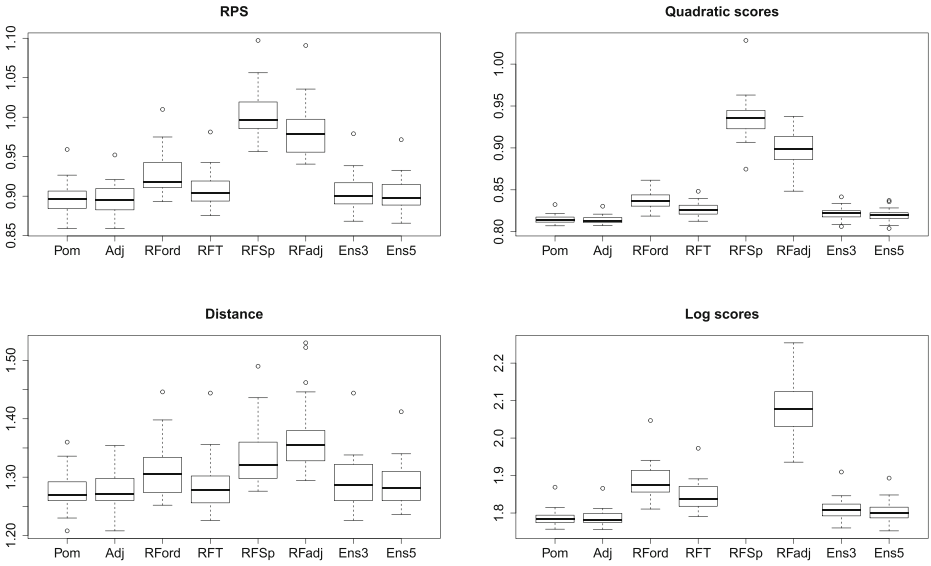

Fig. 10 Results for GLES data ($n_L = 300$, seven response categories)

The best and most stable performance is seen for the ensemble methods that combine parametric and nonparametric methods. Their prediction performance can be considered equivalent to the best method for a particular data set. They seem to efficiently combine the best of two worlds yielding small errors for all data sets. Thus, if one wants to avoid ending

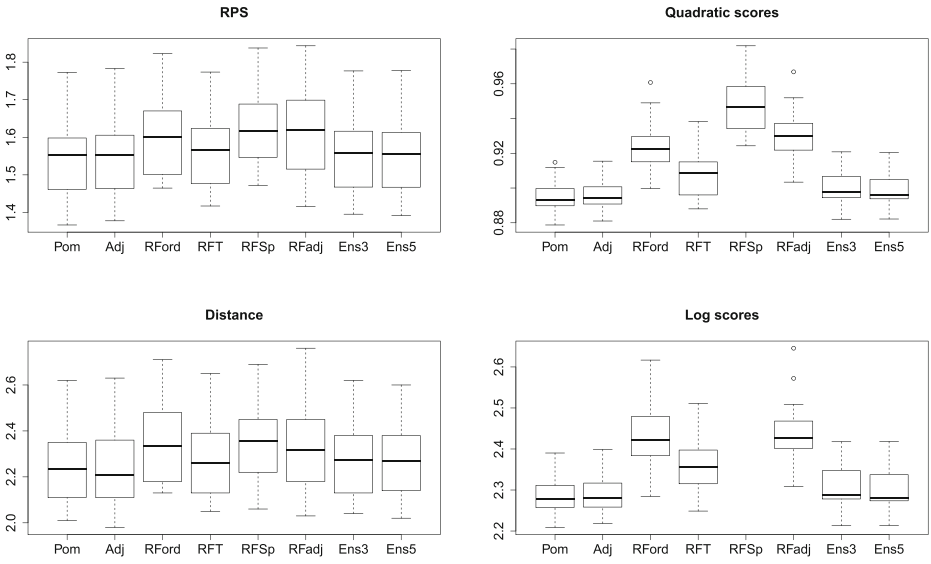

Fig. 11 Results for safety data ($n_L = 400$, ten response categories)

up with an inferior prediction tool one should consider not only trees or parametric models but use a combination of these methods.

For illustration we give the average weights in the ensemble *Ens5* for two data sets. For the housing data, in which the parametric models showed poor performance, the weights for the proportional odds model, the adjacent categories model, ordinalForest fit, and adjacent categories and split-based random forest were 0.146, 0.084, 0.250, 0.279, 0.240. Weights for the parametric models were distinctly smaller than weights for the random forest methods. For the GLES data the corresponding weights were 0.299, 0.303, 0.199, 0.118, and 0.077; thus, the parametric models, which showed better performance, obtained much higher weights than the other methods.

In the evaluations the split-based random forests utilize the *randomForest* method to fit the contained binary trees. Very similar performance is found when using alternative methods to fit binary trees, like *ctree* or *ordinalForest*. These alternative methods yield different trees. When investigating single trees the choice of the method definitely makes a difference, and specific trees may offer advantages, for example conditional trees, which use tests in the splitting procedure, are able to control the significance level and avoid selection bias (Strobl et al., 2007; Hothorn et al., 2006) making them an attractive choice. However, for ensembles of trees as random forests the performance is very similar, at least in the case of split-based based and adjacent categories forests.

5 Importance of Variables

While single trees for split variables are easy to interpret this does not hold for ensembles of trees. Since variables appear in different trees at different positions the impact of variables is hard to infer from plots of hundreds of trees. On the other hand random forests allow for complex effects of predictors, which makes it a flexible prediction tool.

There is a considerable amount of literature that deals with the development of importance measures for random forests (see, for example, Strobl et al., 2007; Strobl et al., 2008; Hapfelmeier et al., 2014; Gregorutti et al., 2017; Hothorn & Zeileis, 2015). A naive measure simply counts the number of times each variable is selected by the individual trees in the ensemble. Better, more elaborate variable importance measures incorporate a (weighted) mean of the individual trees' improvement in the splitting criterion produced by each variable. An example for such a measure is the "Gini importance" available in the randomForest package. It describes the improvement in the "Gini gain" splitting criterion. Alternative, and better variable importance measures are based on permutations yielding so-called permutation accuracy importance measures (Strobl et al., 2007). By randomly permuting single predictor variables X_j, the original association with the response Y is broken. When the permuted variable X_j, together with the remaining un-permuted predictor variables, is used to predict the response, the prediction accuracy is supposed to decrease if the variable X_j had an additional impact on explaining the response. The difference in prediction accuracy before and after permuting X_j yields a permutation accuracy importance measure.

In the following we use the heart data to illustrate how importance measures can be obtained for split-based and adjacent categories random forests. Of course it depends on the algorithm that is used to grow binary trees which importance measure can be computed. Figure 12 shows the Gini importance when using randomForest to fit the binary random

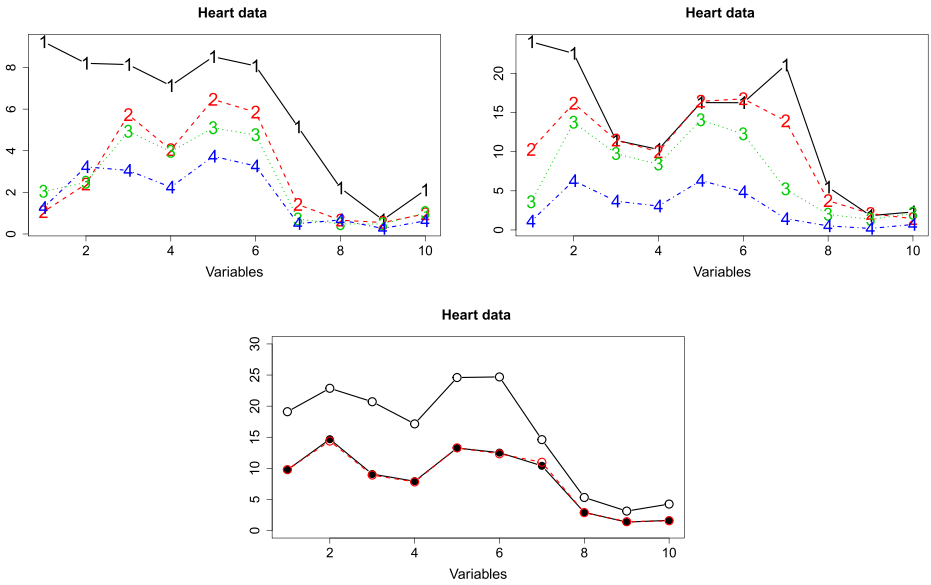

Fig. 12 Gini importance for heart data; variables 1 to 10: chest pain, oldpeak, age, trestbps, chol, thalach, exang, sex, fbs, restecg (randomForest fit); left upper panel: importance for conditional splits in adjacent categories RF, right upper panel: importance for splits in split-based RF, lower panel: averaged importance measures for splits in adjacent categories and split-based RF (lower curves) and multi-categorical fit of randomForest (upper curve)

forests. In the upper panels one sees the importance measures obtained for the split variables, that is, for conditional splits in adjacent categories RF on the left, and direct splits for split-based RF on the right. The numbers 1 to 4 indicate the splits. For example, 3 means that the split is between categories $\{1, 2, 3\}$ and $\{4\}$. It is seen that the first six variables show strong importance with the importance being stronger for lower categories splits and weaker for higher category splits. The lower panel shows the importance measures averaged across the splits. The lower curves, which are almost identical, show the average for the adjacent categories and split-based random forest. It shows, in addition, the Gini importance for the multi-category random forest obtained from randomForest. It is seen that the importance measures have the same order for all the fitted random forests. That the values of importance for the multi-category random forest is higher than for the other two forests is merely a scaling effect.

Figure 13 shows the corresponding picture if conditional trees (cforest) are used , which compare binary predictions before and after permuting. Conditional trees avoid the bias that is found if categorical variables with varying numbers of categories and a mixture of categorical and continuous predictors are used (see, for example, Strobl et al., 2007). Consequently, the obtained importance measures differ from the Gini importance measures. It is seen that variables 1, 2 and 7 are very influential. In particular the importance of variable 1, which is a categorical variable, is more distinct than in Gini importance measures.

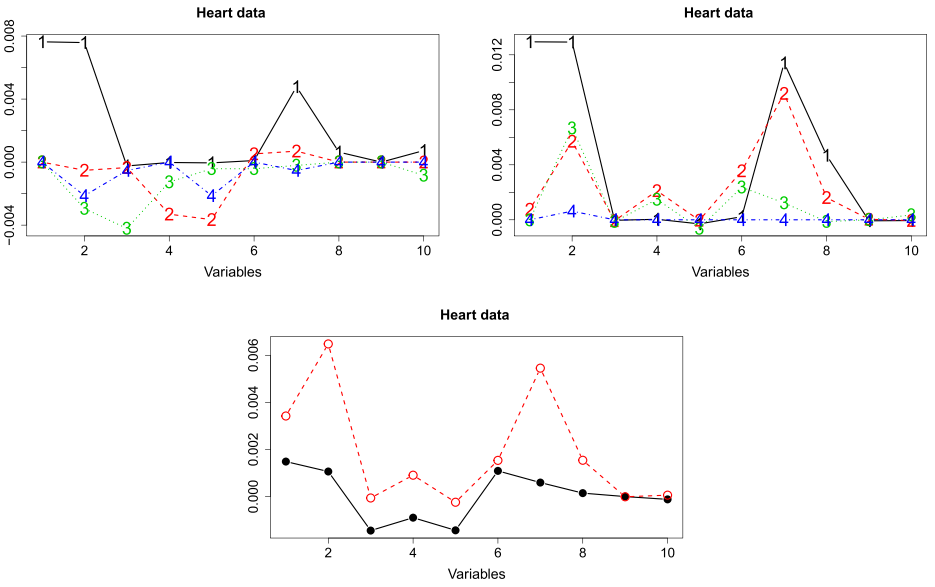

Fig. 13 Importance for heart data; variables 1 to 10: chest pain, oldpeak, age, trestbps, chol, thalach, exang, sex, fbs, restecg (cforest fit); left upper panel: importance for conditional splits in adjacent categories RF, right upper panel: importance for splits in split-based RF, lower panel: averaged importance measures for splits in adjacent categories (lower curve) and split-based RF (upper curve)

6 Concluding Remarks

The split variables, which are the building blocks of ordinal models, have been used to develop ordinal trees and random forests. The basic concept can also be used to generate alternative parametric or nonparametric classification methods that account for the order in responses. One can, for example, use two-class linear discriminant analysis or binary models with variables selection by lasso in the case of many predictors, or use nonparametric methods as the nearest neighborhood classifier for two classes. All of these methods can be used to model the split variables conditionally or unconditionally. In the present paper we restricted consideration to random forests since the objective was to construct score-free random forests.

Also the more recently proposed ordinal random forests are in some way inspired by parametric ordinal models but in a different way than the split variables approach propagated here. The score-free random forests proposed by Buri and Hothorn (2020) follow a quite different strategy to obtain random forests. They fit a cumulative logit model and use the likelihood contributions of the observations to obtain test statistics. The core idea is to regress the obtained partial derivatives of the log-likelihood on prognostic variables. By using the cumulative model the order of categories is used without the need for assigned scores. But it should be noted that the "pure" cumulative model is fitted in subpopulations without including predictors. The ordinal forest propagated by Hornung (2020) also uses the cumulative logistic model. It exploits the latent continuous response variable underlying the observed ordinal response variable by explicitly using the widths of the adjacent intervals in the range of the continuous response variable. These intervals are considered as corresponding to the classes of the ordinal response variable. That means, "the ordinal

response variable is treated as a continuous variable, where the differing extents of the individual classes of the ordinal response variable are implicitly taken into account" (Hornung, 2020). The approach is closely related to conventional random forests for continuous outcomes but optimizes the assigned scores instead of considering them as given, and therefore is score-free in a certain sense.

The accuracy measures obtained for the data sets suggest that one might distinguish between three types of data sets data, data for which the parametric models perform distinctly better, data sets where there is not much difference between approaches, and data sets, for which parametric models clearly outperform random forests. In particular the latter type of data suggests that one should not rely on random forests to always perform well. Split-based random forests seem to compete well with the ordinal forests that have proposed recently in the literature only in cases where it is sensible to use random forests since they are stronger than simple parametric models. In cases where there is not much to gain from using random forests all the random forests approaches tend to have comparable performance. If parametric models clearly outperform random forests the best random forests methods is the RFT method. The most stable performance can be expected from the ensemble methods, which tend to perform at least as well as the best method.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of Interest The author declares no competing interests.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*, 2nd edn. Wiley: New York.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society B*, *46*, 1–30.

Anderson, J. A., & Phillips, R. R. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Applied Statistics*, *30*, 22–31.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any 'threshold disorder controversy'. *Educational and Psychological Measurement*, *73*(1), 78–124.

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

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non–proportional odds. *Journal of Clinical Epidemiology*, *51*, 809–816.

Biernacki, C., & Jacques, J. (2016). Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing*, *26*(5), 929–943.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, *46*, 1171–1178.
Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.
Bühlmann, P., Yu, B., et al. (2002). Analyzing bagging. *The Annals of Statistics*, *30*(4), 927–961.
Buri, M., & Hothorn, T. (2020). Model-based random forests for ordinal regression. The International Journal of Biostatistics 1(ahead-of-print).
Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information–theoretic approach*. New York: Springer.
Campbell, M. K., & Donner, A. P. (1989). Classification efficiency of multinomial logistic-regression relative to ordinal logistic-regression. *Journal of the American Statistical Association*, *84*(406), 587–591.
Campbell, M. K., Donner, A. P., & Webster, K.M. (1991). Are ordinal models useful for classification? *Statistics in Medicine*, *10*, 383–394.
Cappelli, C., Simone, R., & Di Iorio F. (2019). cubremot: a tool for building model-based trees for ordinal responses. *Expert Systems with Applications*, *124*, 39–49.
Chernozhukov, V., Fernández-Val, I., & Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, *81*(6), 2205–2268.
Chu, W., & Keerthi, S. S. (2007). Support vector ordinal regression. *Neural Computation*, *19*(3), 792–815.
Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553.
Cox, C. (1995). Location-scale cumulative odds models for ordinal data: a generalized non-linear model approach. *Statistics in Medicine*, *14*, 1191–1203.
Deb, P., & Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, *12*(3), 313–336.
Fernandez, D., Liu, I., & Costilla, R. (2019). A method for ordinal outcomes: the ordered stereotype model. *International Journal of Methods in Psychiatric Research*, *28*, e1801.
Foresi, S., & Peracchi, F. (1995). The conditional distribution of excess returns: an empirical analysis. *Journal of the American Statistical Association*, *90*(430), 451–466.
Galimberti, G., Soffritti, G., & Di Maso, M. (2012). Classification trees for ordinal responses in r: the rpartscore package. *Journal of Statistical Software, 47*.
Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–376.
Goodman, L. A. (1981a). Association models and canonical correlation in the analysis of cross-classification having ordered categories. *Journal of the American Statistical Association*, *76*, 320–334.
Goodman, L. A. (1981b). Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika*, *68*, 347–355.
Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, *13*, 1665–1677.
Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678.
Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, *24*(1), 21–34.
Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102.
Hornung, R. (2020). Ordinal forests. *Journal of Classification*, *37*, 4–17.
Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651–674.
Hothorn, T., Lausen, B., Benner, A., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in Medicine*, *23*(1), 77–91.
Hothorn, T., & Zeileis, A. (2015). partykit: a modular toolkit for recursive partytioning in r. *The Journal of Machine Learning Research*, *16*(1), 3905–3909.
Iannario, M., Piccolo, D., & Simone, R. (2020). CUB: a class of mixture models for ordinal data. R package version 1.1.4. http://cran.r-project.org/package=cub.
Janitza, S., Tutz, G., & Boulesteix, A.-L. (2016). Random forest for ordinal responses: prediction and variable selection. *Computational Statistics & Data Analysis*, *96*, 57–73.
Kateri, M. (2014). *Contingency table analysis*. Berlin: Springer.
Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2020). Ensemble of optimal trees, random forest and random projection ensemble classification. *Advances in Data Analysis and Classification*, *14*(1), 97–116.

Kim, J.-H. (2003). Assessing practical significance of the proportional odds assumption. *Statistics & probability letters*, *65*(3), 233–239.
Kleiber, C., & Zeileis, A. (2008). *Applied Econometrics with R*. Springer: New York.
Liaw, A., Wiener, M., Breiman, L., & Cutler, A. (2015). Package randomforest.
Liu, I., Mukherjee, B., Suesse, T., Sparrow, D., & Park, S.K. (2009). Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statistics in Medicine*, *28*(3), 412–429.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.
Masters, G. N., & Wright, B. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*, 529–544.
McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B*, *42*, 109–127.
Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(Jun), 983–999.
Muraki, E. (1997). A generalized partial credit model. *Handbook of modern item response theory*, pp 153–164.
Peterson, B., & Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, *39*, 205–217.
Piccolo, D., & Simone, R. (2019). The class of CUB models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications*, *28*(3), 389–435.
Polikar, R. (2009). Ensemble learning. *Scholarpedia*, *4*(1), 2776.
Rattinger, H., Roßteutscher, S., Schmitt-beck, R., Weßels, B., & Wolf, C. (2014). Pre-election cross section (GLES 2013). *GESIS Data Archive, Cologne ZA5700 Data file Version 2.0.0*.
Rudolfer, S. M., Watson, P. C., & Lesaffre, E. (1995). Are ordinal models useful for classification? A revised analysis. *Journal of Statistical Computation Simulation*, *52*(2), 105–132.
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.
Simone, R., & Tutz, G. (2020). Hybrid random forests for ordinal data. In N. Salvati, A. Pollice, & F. Schirripa Spagnolo (Eds.) *Book of short papers SIS* (pp. 1171–1176).
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307.
Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25.
Tutz, G. (2012). *Regression for categorical data*. Cambridge University Press.
Tutz, G. (2020). Ordinal regression: a review and a taxonomy of models. *Wiley Interdisciplinary Reviews: Computational Statistics*, pp e1545.
Ursino, M., & 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.
Winham, S. J., Freimuth, R. R., & Biernacka, J.M. (2013). A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: the ASA Data Science Journal*, *6*(6), 496–505.