Generalised Linear Model Trees with Global Additive Effects

Seibold, Heidi ; Hothorn, Torsten ; Zeileis, Achim

Generalised Linear Model Trees with Global Additive Effects

Heidi Seibold
University of Zurich

Torsten Hothorn
University of Zurich

Achim Zeileis
Universität Innsbruck

Abstract

Model-based trees are used to find subgroups in data which differ with respect to model parameters. In some applications it is natural to keep some parameters fixed globally for all observations while asking if and how other parameters vary across the subgroups. Existing implementations of model-based trees can only deal with the scenario where all parameters depend on the subgroups. We propose partially additive linear model trees (PALM trees) as an extention to (generalised) linear model trees (LM and GLM trees, respectively), in which the model parameters are specified a priori to be estimated either globally from all observations or locally from the observations within the subgroups determined by the tree. Simulations show that the method has high power for detection of subgroups in the presence of global effects and reliably recovers the true parameters. Furthermore, treatment-subgroup differences are detected in an empirical application of the method to data from a mathematics exam: the PALM tree is able to detect a small subgroup of students that had a disadvantage in an exam with two versions while adjusting for overall ability effects.

Keywords: subgroup analysis, model-based recursive partitioning, GLM, tree.

1. Introduction

Model-based recursive partitioning (Zeileis, Hothorn, and Hornik 2008) is used to partition data into groups that differ in terms of the parameters in the model. Less technically it finds subgroups in a clinical trial which differ in terms of treatment effect on a health score or areas in a city which differ in terms of the influence of square metres on the rent price. Sometimes there are parameters in the model that one wants to fix for all groups, e.g. the effect of smoking on the health outcome in the clinical trial or the effect of inflation/deflation on rent prices. This, however, is not possible in model-based recursive partitioning as described in Zeileis *et al.* (2008). Here we propose an algorithm called PALM tree that is similar to model-based recursive partitioning but allows fixing parameters over all groups, i.e. only some parameters depend on the tree structure.

There have been several developments in the past years toward the direction of combining models and trees, where one part of the model follows a tree structure and one part does not. The Simultaneous Threshold Interaction Modeling Algorithm (STIMA, Dusseldorp, Conversano, and Van Os 2010) starts off with a main effects model and adds interactions based on a tree. Fokkema, Smits, Zeileis, Hothorn, and Kelderman (2015) proposed GLMM tree, a method that is similar to PALM tree, but is used to keep random effects in a generalised

linear mixed-effects model (GLMM) fixed instead of – as in PALM tree – further fixed effects. Other approaches going in the direction of GLMM tree are RE-EM tree (Sela and Simonoff 2012) and MERT (Hajjem, Bellavance, and Larocque 2011).

In the literature on subgroup analyses in the estimation of treatment effects, special tree-based procedures have been proposed. These methods are commonly used in the analysis of clinical trials, but are equally relevant in contexts such as marketing studies evaluating different marketing strategies or studies on website user behaviour, where users are randomly served one of two website versions (A/B testing). Sies and Van Mechelen (2016) review some of the methods in a setting where there are covariates that should be fixed. One promising method this review is a method by Zhang, Tsiatis, Davidian, Zhang, and Laber (2012) which estimates rules of optimal treatment for each patient subgroup (optimal treatment regimes).

In Section 2.1 we will describe how PALM trees are computed and show their connection to (generalised) linear models and model-based recursive partitioning, in particular LM trees and GLM trees. Furthermore we will show how model-based trees (LM trees, GLM trees and PALM trees) can be used in finding subgroups with differential treatment effects. In section 3 we show the results of a simulation study in which we compare LM tree, PALM tree and the optimal treatment regime method by Zhang *et al.* (2012). In section 4 we will apply the PALM tree to data of a mathematics exam, where the endpoint is performance in the exam, the "treatment" is the student group (early morning or late group) and the known prognostic factor is the performance in online tests the students participate in during the semester. Finally we will discuss strengths and limitations of model-based trees in general and PALM trees in particular.

2. Methods

2.1. PALM tree and (G)LM tree

Going from (generalised) linear models – (G)LMs – via (G)LM trees to PALM trees can be viewed as an evolutionary process where one method evolves from the other. In the following we show how to obtain a GLM tree or PALM tree by partitioning based on a GLM (we are focusing on GLMs since LMs are merely a special case of GLMs).

Methodology

The goal of GLMs, GLM trees and PALM trees is to appropriately estimate the effect of covariates $\mathbf{x}$ on an outcome $\mathbf{y}$. The main difference between the tree methods is the structure of the linear predictor. While a GLM contains linear effects $\boldsymbol{\beta}$, a GLM tree contains linear effects $\boldsymbol{\beta}(\mathbf{z})$ within each subgroup. These subgroups are defined by variables $\mathbf{z}$. A PALM tree contains globally *fixed* linear effects $\boldsymbol{\gamma}$ for some covariates $\mathbf{x}_F$ and subgroup-wise *varying* linear effects $\boldsymbol{\beta}(\mathbf{z})$ for other covariates $\mathbf{x}_V$. Mathematically this can be expressed as follows:

$$\text{GLM} \quad g(\boldsymbol{\mu}) = \mathbf{x}^\top \boldsymbol{\beta} \tag{1}$$

$$\text{GLM tree} \quad g(\boldsymbol{\mu}) = \mathbf{x}^\top \boldsymbol{\beta}(\mathbf{z}) \tag{2}$$

$$\text{PALM tree} \quad g(\boldsymbol{\mu}) = \mathbf{x}_V^\top \boldsymbol{\beta}(\mathbf{z}) + \mathbf{x}_F^\top \boldsymbol{\gamma} \tag{3}$$

with expected response $\boldsymbol{\mu} = \mathbb{E}(\mathbf{y})$ and link function g. $\boldsymbol{\beta}(\mathbf{z})$ is the interaction effect between

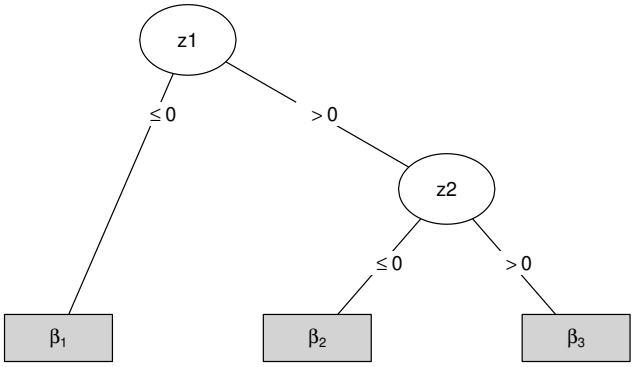


Figure 1: Example of a model-based tree.

covariates $\mathbf{x}$ (or $\mathbf{x}_V$) and the subgroups (defined by variables $\mathbf{z}$). If the subgroup structure is known, models (2) and (3) are just GLMs with interactions, i.e. Equation (1) = Equation (2) = Equation (3). Figure 1 shows an example of a subgroup structure with $\boldsymbol{\beta}(\mathbf{z})$ defined as

$$\boldsymbol{\beta}(\mathbf{z}) = \begin{cases} \boldsymbol{\beta}_1 & \text{if } \; z_1 \leq 0 \\ \boldsymbol{\beta}_2 & \text{if } \; (z_1 > 0) \; \wedge \; (z_2 \leq 0) \\ \boldsymbol{\beta}_3 & \text{if } \; (z_1 > 0) \; \wedge \; (z_2 > 0). \end{cases} \tag{4}$$

where $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_3$. If the three subgroups are known, only the parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ have to be estimated, i.e.

$$\begin{aligned} \mathbf{x}^\top \boldsymbol{\beta}(\mathbf{z}) = \; & I(z_1 \leq 0) \cdot \mathbf{x}^\top \boldsymbol{\beta}_1 + \\ & I((z_1 > 0) \; \wedge \; (z_2 \leq 0)) \cdot \mathbf{x}^\top \boldsymbol{\beta}_2 + \\ & I((z_1 > 0) \; \wedge \; (z_2 > 0)) \cdot \mathbf{x}^\top \boldsymbol{\beta}_3 \\ = \; & \tilde{\mathbf{x}}^\top \tilde{\boldsymbol{\beta}}. \end{aligned} \tag{5}$$

However, the subgroup structure is usually unknown and a GLM tree is needed to estimate the tree and the corresponding parameters $\boldsymbol{\beta}(\mathbf{z})$ simultaneously.

GLM trees assume that all parameters have to be subgroup specific. This does not necessarily have to be the case. So PALM trees offer a compromise between GLM trees and GLMs by having one part in which the parameters depend on subgroups and another part in which the parameters are the same for all subjects / subgroups. The global parameter vector $\boldsymbol{\gamma}$ is the same for all subgroups. If $\boldsymbol{\gamma}$ is known, $\mathbf{x}_F^\top \boldsymbol{\gamma}$ can be included in the model as an offset and the estimation of models (2) and (3) are the same. However, also $\boldsymbol{\gamma}$ is usually also unknown and a PALM tree is needed to estimate both the subgroup structure and the parameters $\boldsymbol{\beta}(\mathbf{z})$ and $\boldsymbol{\gamma}$. Note that variables $\mathbf{x}_F$ with a global effect $\boldsymbol{\gamma}$ have to be defined a priori and usually $\mathbf{x}_V$ and $\mathbf{x}_F$ do not overlap (other than $\mathbf{x}$ and $\mathbf{z}$ where overlap is possible).

Algorithm

We now describe the detailed algorithms of GLM trees and PALM trees, starting with GLM trees. The GLM tree algorithm is not new and has been explained in depth by Zeileis *et al.* (2008). The following description of the algorithm focuses on the parts that are necessary in order to demonstrate the full concept of the PALM tree algorithm. GLM trees are grown as follows, starting with the root node containing all observations:

1. Compute model (1), or equivalently model (2) with a single subgroup ($\boldsymbol{\beta}(\mathbf{z}) = \boldsymbol{\beta}$), in the given node.
2. Test for instability in the model parameters with respect to each of the possible subgroup defining variables $\mathbf{z}_1, \ldots, \mathbf{z}_J$ (using correction for multiple testing).
3. If overall test is significant, choose the $\mathbf{z}_j$ corresponding to the lowest p-value as the split variable.
4. Choose split point as point which maximises the sum of the likelihoods in the two emerging groups.
5. Iterate steps 1 to 4 until test is not significant or other stop criterion is fulfilled.

The result is groups which differ with respect to at least one of the model parameters $\boldsymbol{\beta}$. In practice, however, all parameters vary slightly between subgroups due to the refitting of the model in each node, i.e. for each group of observed subjects. If there are covariates which in reality influence the response linearly (for all observations), this leads to an overly complex model. The PALM tree algorithm eliminates this downside by introducing the possibility to build models where some parameters are kept stable over subgroups. This is achieved by using an EM-type algorithm that starts off with model 3 with a single subgroup, i.e. $\boldsymbol{\beta}(\mathbf{z}) = \boldsymbol{\beta}$, and iterates between

- estimating $\boldsymbol{\gamma}$ for a given tree structure and
- estimating the tree structure for a given $\hat{\boldsymbol{\gamma}}$.

By using this iterative process we ensure that in each step only one is unknown:Either $\boldsymbol{\gamma}$ or the tree structure. If the tree structure is known we can estimate model 3 using the known subgroup $\times$ covariate ($\mathbf{x}_V$) interactions. If $\hat{\boldsymbol{\gamma}}$ is known it can be included in the model as an offset and the tree can be estimated as a GLM tree. Note that $\boldsymbol{\beta}(\mathbf{z})$ is estimated in both steps. The algorithm stops when no (or very little) improvement can be achieved. This is usually the case when the tree converges and does not change anymore.

2.2. Special application: Treatment effects

One common application of model-based trees is for subgroup analyses in clinical trials (Lipkovich, Dmitrienko, and D'Agostino 2016). In the simplest case one is interested in a treatment effect of a new treatment versus standard of care or no treatment, i.e. $\mathbf{x}$ or $\mathbf{x}_V = (1, \mathbf{x}_A)$ with $x_{Ai} = I(\text{patient } i \text{ received new treatment})$. In this setting one differentiates between prognostic and predictive factors (Italiano 2011). Prognostic factors are patient characteristics (measured before treatment start) which directly impact the response, e.g. a health score. Predictive factors are patient characteristics which impact the efficacy of the treatment. In

the PALM tree framework, predictive factors should be included in the split variables $\mathbf{z}$ and prognostic factors, if known in advance, can be included in $\mathbf{x}_F$. Another term is frequently used in subgroup analyses for treatment effects: treatment regimes or optimal treatment regimes. An optimal treatment regime is a rule which indicates which treatment is better in which subgroup. Treatment regimes only check the sign of the treatment effect in each subgroup. If they differ between subgroups, the treatment effects are called qualitative; if one treatment is better than the other in all subgroups, they are called quantitative.

2.3. Comparison to other approaches

GLMM tree (Fokkema *et al.* 2015) is a method closely related to PALM tree, which also builds on the GLM tree algorithm and like PALM tree keeps parts of the model stable. The major difference is the fact that GLMM trees focus, as the name says, on generalised mixed effects models and the part that is being kept stable across subgroups are the random effects.

STIMA (Dusseldorp *et al.* 2010) is a tree algorithm where the first split is made in an a priori specified variable, which in the treatment case is the treatment indicator. All further splits are found by an exhaustive search and finally a cross-validation based pruning procedure is run to find the optimal tree. STIMA is similar to PALM tree in the sense that it starts off with a main effects model and new splits are selected based on a measure of variance-accounted-for. The main effects of the model are kept stable across groups and additional effects are added to the model based on the tree structure. A very similar approach is called partially linear tree-based regression model (PLTR, Chen, Yu, Hsing, and Therneau 2007; Mbogning and Toussile 2015), which was initially invented to analyse gene-gene and gene-environment effects.

The approach by Zhang *et al.* (2012) aims to estimate optimal treatment regimes and is only used in the treatment effect application. In the following we will use the term OTR (optimal treatment regimes) for this method. OTR is not as closely related to PALM tree as the previously mentioned methods, but has shown good performance in settings in which PALM trees are appropriate (Sies and Van Mechelen 2016). OTR does not target estimating the treatment effect itself but targets learning which treatment is superior for certain groups of patients. The algorithm starts off with the so called outcome model, which includes main effects and treatment covariate interactions. After estimating the model the algorithm proceeds as follows:

1. For all patients in the training data predict the response under treatment $\hat{\boldsymbol{\mu}}_1$ and under control $\hat{\boldsymbol{\mu}}_0$ from the outcome model. Determine the difference $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$ between the two.
2. Compute a classification algorithm using $I(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0 > 0)$ as response and $|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0|$ as weights.

Any classification method that can deal with (non-integer) weights could be used in step 2.

3. Simulation study

We compare performance between PALM trees, LM trees and the trees grown based on the algorithm proposed by Zhang *et al.* (2012) in the treatment effect setting. However, this is also relevant to other settings. The aim is to evaluate the methods with respect to (1) finding the

Simulation variable	Default	Variation	# Values
Difference in treatment effects Δ_β	0.5	0.1–1.5	8
Number of observations n	300	100–900	5
Qualitative treatment $\times$ covariate interaction	Yes	Yes/No	2
Number of patient characteristics m	30	10–70	4
Number of predictive factors p	2	1–4, 0	4, 1
Number of prognostic factors q	2	1–4	4

Table 1: Simulation settings. For each scenario one simulation variable is varied and the rest are kept to the standard value. The value $p = 0$ is only used for the assessment of the type 1 error rate (Section 3.2).

correct subgroups (Section 3.1), (2) not splitting when ther are no subgroups (Section 3.2), (3) finding the optimal treatment regime (Section 3.3), and (4) correctly estimating the treatment effect (Section 3.4). Note that evaluations (1) and (2) are connected in the sense that they both evaluate the ability to find the correct subgroups. Furthermore, (3) and (4) are connected in the sense that the both evaluate the ability to give good treatment recommendations.

We simulate a binary variable (treatment indicator) $\mathbf{x}_A$ which is either 1 or 0, each with probability 0.5 and m correlated variables (patient characteristics)

$$\mathbf{Z} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{\Sigma}) \tag{6}$$

with

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & \cdots & 0.2 \\ \vdots & \vdots & \ddots & \vdots \\ 0.2 & 0.2 & \cdots & 1 \end{pmatrix}. \tag{7}$$

We define the first p variables $\mathbf{z}_1, \ldots, \mathbf{z}_p$ to be the true predictive factors, i.e. the patient characteristics that actually interact with the treatment and thus pose relevant split variables. The cutpoint is always at $z_j = 0$ and the subsequent split is always in the subgroup with $z_j > 0$, i.e. on the right side of the tree when visualised as in Figure 1. We define the consecutive q variables $\mathbf{x}_F = (\mathbf{z}_{p+1}, \ldots, \mathbf{z}_{p+q})$ to be the true and known prognostic factors. All further patient characteristics $\mathbf{z}_{p+q+1}, \ldots, \mathbf{z}_m$ are noise variables. We simulate the response variable $\mathbf{y}$ with

$$\mathbf{y} = \mathbf{x}_A^\top \beta(\mathbf{z}) + \mathbf{x}_F^\top \boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{8}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 1.5)$ is the error term. The effect of the prognostic factors is set to $\boldsymbol{\gamma} = \mathbf{1}$. $\beta(\mathbf{z})$ follows a tree structure, which for the scenarios with $p = 2$ is visualised in Figure 1. The mathematical representation is as in Equation (4) with a fixed difference between the effects in the subgroups Δ_β. We define a default simulation scenario, which is shown in the second column of Table 1. In this default scenario $\Delta_\beta = 0.5$ and

$$\beta(\mathbf{z}) = \begin{cases} -0.375 & = \beta_1 & \text{if } z_1 \leq 0 \\ 0.125 & = \beta_2 = \beta_1 + \Delta_\beta & \text{if } z_1 > 0 \ \wedge \ z_2 \leq 0 \\ 0.625 & = \beta_3 = \beta_2 + \Delta_\beta & \text{if } z_1 > 0 \ \wedge \ z_2 > 0. \end{cases} \tag{9}$$

To obtain a diverse set of simulation scenarios which are comparable, we fix all but one of the simulation variables to the default. The range of variation of each simulation variable is given in the third column of Table 1 alongside the number of equidistant values considered (# Values). From this we get all necessary information about the simulation, e.g. q takes 4 different values $1, 2, 3, 4$. For each distinct simulation setting we simulate 150 data sets. Note that just for the assessment of the type 1 error rate (Section 3.2) the number of predictive factors is set to zero. For the simulation scenarios where $p \neq 2$ and thus less/more than three true subgroups exist, $\beta(\mathbf{z})$ follows the same logic as in Equation (9), i.e. $\beta_b = \beta_{b-1} + \Delta_\beta$ for $b = 2, \ldots, (p+1)$. The value of β_1 depends on whether the first split is qualitative or not and on Δ_β. If the first split is not qualitative then $\beta(1) = 0.5$. If the first split is qualitative $\beta(1) = -3/4 \cdot \Delta_\beta$. This also means that any consecutive splits after the first are quantitative. Using the simulated data we compare the following methods:

PALM tree with $\mathbf{x}_V = (\mathbf{1}, \mathbf{x}_A)$ and $\mathbf{x}_F = (z_{p+1}, \ldots, z_{p+q})$. The only way we could have specified this algorithm better for the given data generating process would have been to add the intercept to $\mathbf{x}_F$, but in real application one would usually allow the intercept to vary to account for unknown prognostic factors contained in $\mathbf{z}$.

LM tree 1 with $\mathbf{x} = (\mathbf{1}, \mathbf{x}_A)$. This algorithm is of interest to see how well a misspecified model-based tree behaves. LM tree 1 has to approximate $\mathbf{x}_F^\top \boldsymbol{\gamma}$ using step functions and thus cannot give good results in terms of most measures used below. However, we are interested in how well it can do in terms of estimating the correct treatment regime.

LM tree 2 with $\mathbf{x} = (\mathbf{1}, \mathbf{x}_A, \mathbf{x}_F)$. This tree is expected to behave better than LM tree 1, since it contains the correct covariates in the model, but worse than PALM tree since it may split with respect to instabilities in the parameters for $\mathbf{x}_F$ plus it is overly complex due to the fitting of separate $\mathbf{x}_F$-parameters in each subgroup.

OTR with outcome model $g(\boldsymbol{\mu}) = (\mathbf{1}, \mathbf{x}_A, \mathbf{x}_F)^\top \boldsymbol{\gamma} + (\mathbf{x}_A \circ \mathbf{z})^\top \boldsymbol{\beta}$ (with $\mathbf{x}_A \circ \mathbf{z}$ interaction between $\mathbf{x}_A$ and $\mathbf{z}$) and pruned CARTs (Classification and Regression Trees, Breiman, Friedman, Stone, and Olshen 1984) as classification method. OTR was invented to find optimal treatment regimes and thus is expected to be good at finding the right treatment. OTR is not intended to find quantitative interactions and thus can not be good at this.

3.1. Are the correct subgroups found?

To investigate whether the correct subgroups are captured by the different methods, we look at the number of subgroups found as well as the adjusted rand index (ARI, Hubert and Arabie 1985; Milligan and Cooper 1986). The ARI measures how well the retrieved subgroups fit with the true underlying subgroups. If the subgroups found are similar to the true subgroups the ARI will have a value up to 1. If the subgroups are only as good as a random group assignment the ARI is 0. If there is systematic missclassification, the ARI can also be negative.

The first row of Figure 2 shows the mean number of selected subgroups over the 150 simulated data sets and their corresponding trees for differing *distances between treatment effects* Δ_β and differing *numbers of observations* n. This means we are looking at the case where all variables are kept at the standard value except Δ_β or n respectively. The second row shows the corresponding ARI. The similarity between the PALM tree and LM tree 2 algorithms is

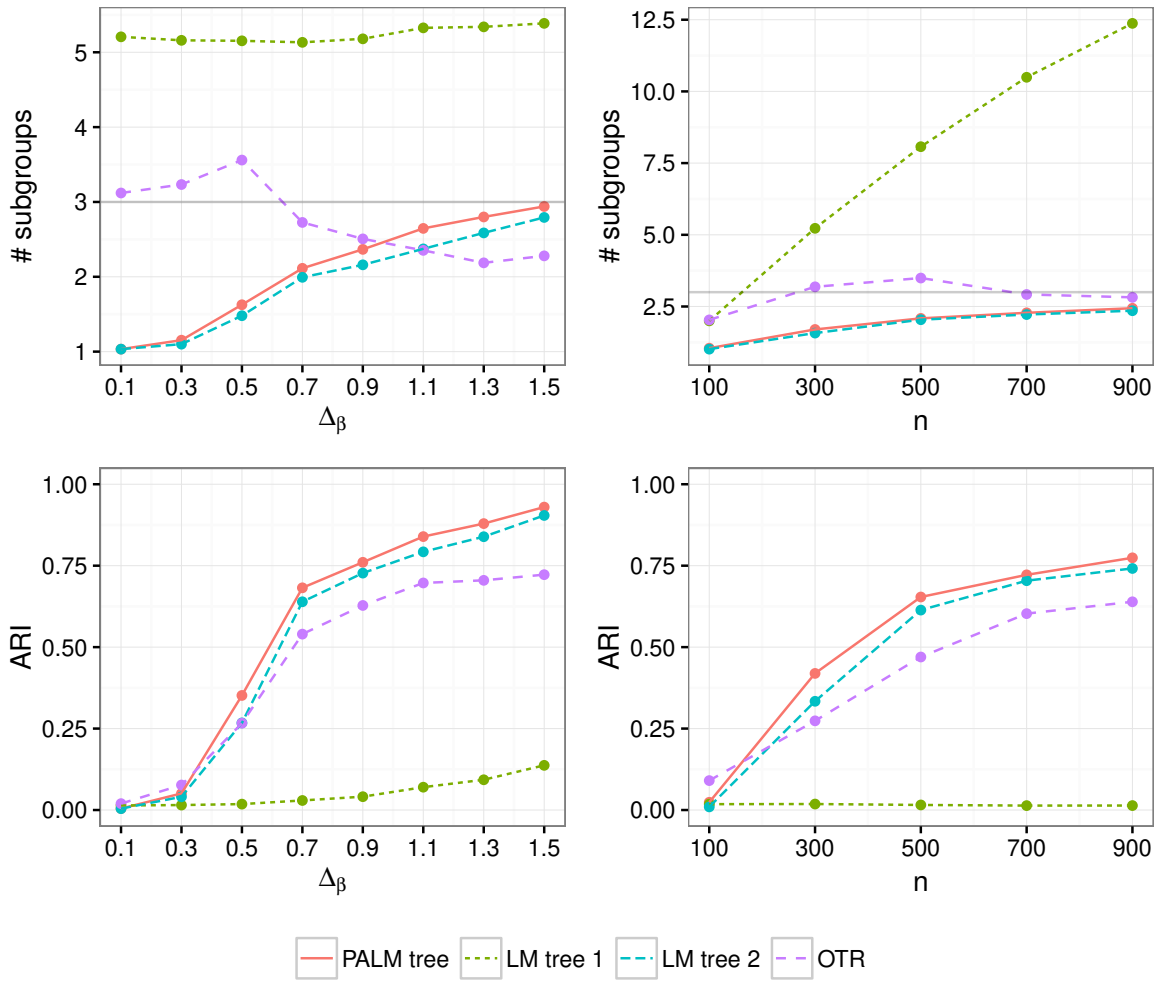


Figure 2: Mean number of subgroups and mean ARI for varying Δ_β and number of observations (Question 3.1).

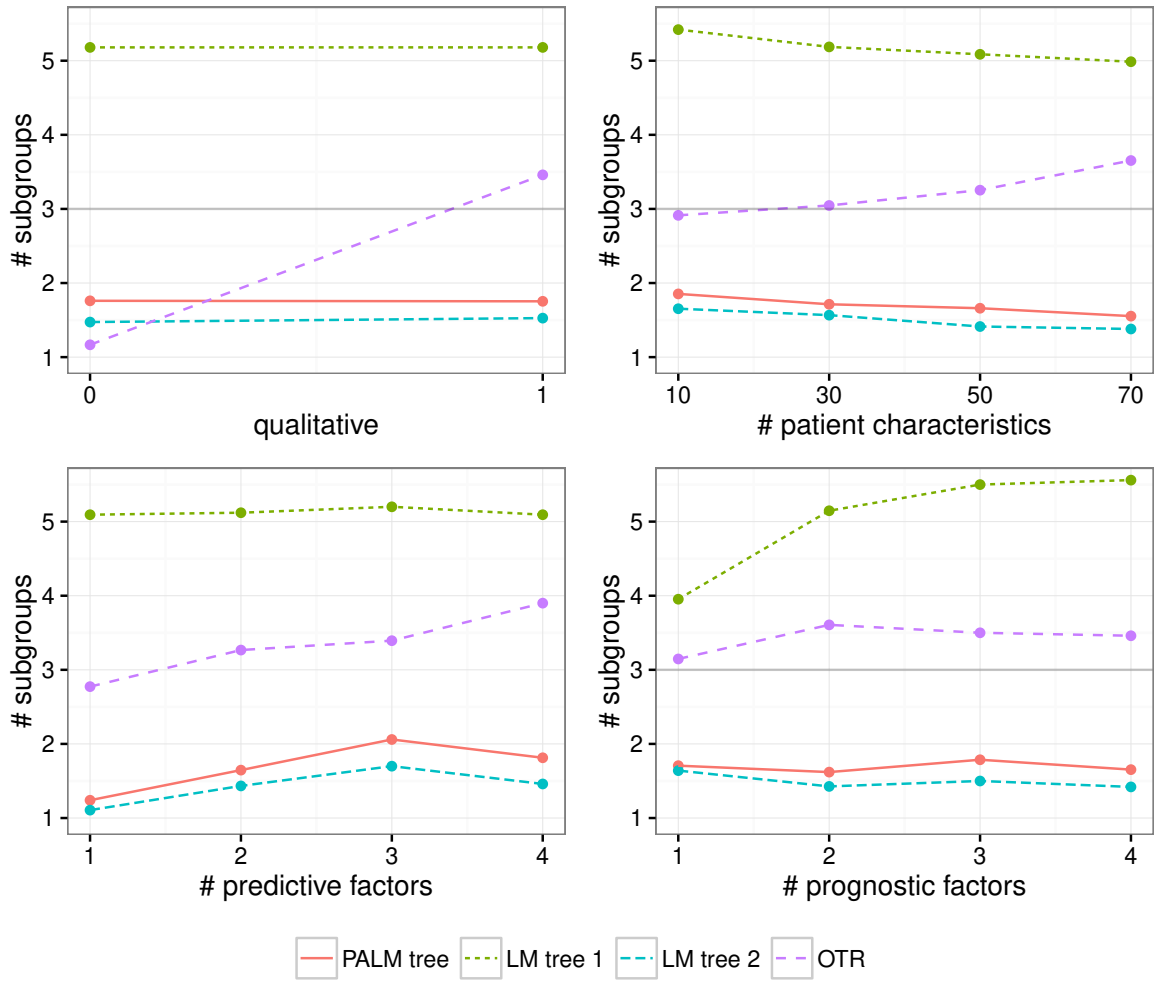


Figure 3: Mean number of subgroups for varying types of subgroups (quantitative/qualitative), number of patient characteristics, predictive factors and prognostic factors (Question 3.1).

obvious. For both the number of subgroups and the ARI the results are very similar, although PALM tree is slightly better. Both algorithms get steadily closer to the optimal solution with increasing Δ_β as well as with increasing number of observations. LM tree 1 performs badly since it approximates the linear relation between the prognostic factors and the response with splits in the data. This is also the reason why with increasing n the number of subgroups increases. This effect muffles the grouping with respect to the treatment effect, even if it gets less with increasing Δ_β. The number of subgroups found for OTR is close to the actual number of subgroups (3 for the given scenarios in Figure 2). However these subgroups actually capture the true subgroups less well than the subgroups from PALM tree and LM tree 2. The ARI for OTR is lower than the ARI of PALM tree and LM tree 2 except for very low values of Δ_β and n, which can be explained by the fact that the model-based trees use statistical tests and CART does not.

Figure 3 shows the mean number of subgroups for the remaining simulation scenarios. The model-based trees are not affected by the *type of subgroup*. OTR, however, is designed to find only qualitative subgroups and thus most of the time finds only one group when there are only quantitatively differing subgroups. For increasing *number of patient characteristics* $\mathbf{z}$, the model-based trees become more conservative and find slightly less subgroups, which is due to the correction for multiple testing (Bonferroni correction). OTR finds more subgroups when more patient characteristics are available. With increasing *number of predictive factors* the number of subgroups should increase. The true number of subgroups is always the number of predictive factors + 1. The lower left panel of Figure 3 shows that only for OTR the number of subgroups increases constantly. This is surprising because only the the first split is qualitative. For the other algorithms the way of how we simulated the data seems to have an impact. With an increasing number of predictive factors the subgroups get smaller and the tests have less power. The only algorithm that is affected by the *number of prognostic factors* is LM tree 1, which corresponds to the fact that there are more linear terms to approximate through the tree structure.

3.2. How often are subgroups found even though there are none?

To investigate the type 1 error rate, i.e. the probability that subgroups are found even though there are none, we simulated data as above, but with no predictive factors. This means the treatment effect is the same for all patients. Figure 4 shows the behaviour of the methods with changing *number of observations*. LM tree 1 has a constant value of 1 here since it finds subgroups that have to do with the prognostic factors. PALM tree performs best, but is conservative for low and high numbers of observations. OTR performs poorly for few observations but improves with more.

3.3. Is the correct treatment predicted to be better?

The next measure we want to look at is the proportion of patients for which the better treatment is correctly identified. This is what OTR was designed to be good at and especially due to the way we simulate data (with a simple interaction) OTR can be expected to perform well. Figure 5 shows the proportion of patients for which the better treatment is correctly identified for the scenarios with varying difference between treatment effects Δ_β and varying number of predictive factors. When the *difference between treatment effects* Δ_β is small it is difficult for all methods to predict the correct treatment regime. For $\Delta_\beta = 0.1$ it is close to

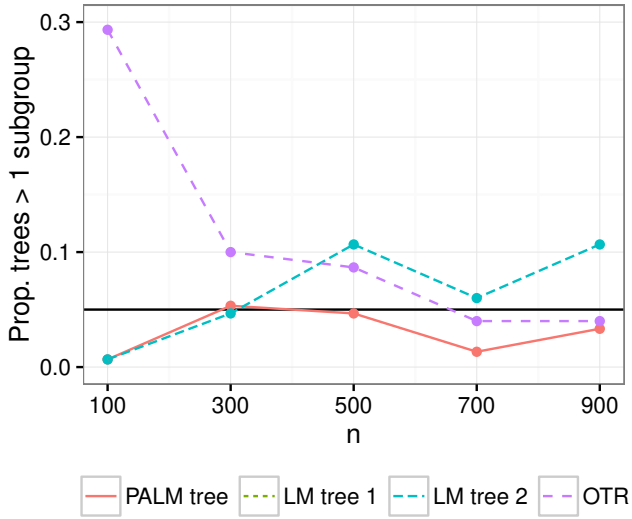


Figure 4: Proportion of trees with more than one subgroup for varying number of observations (Question 3.2). Black line at 0.05.

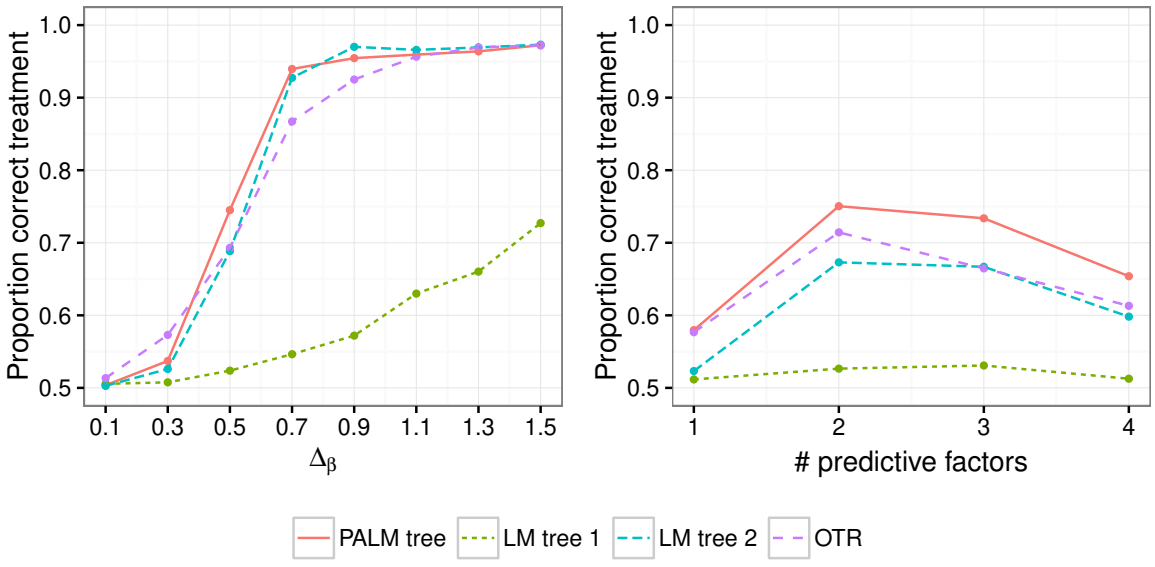


Figure 5: Proportion of observations in all trees where better treatment is correctly identified (Question 3.3).

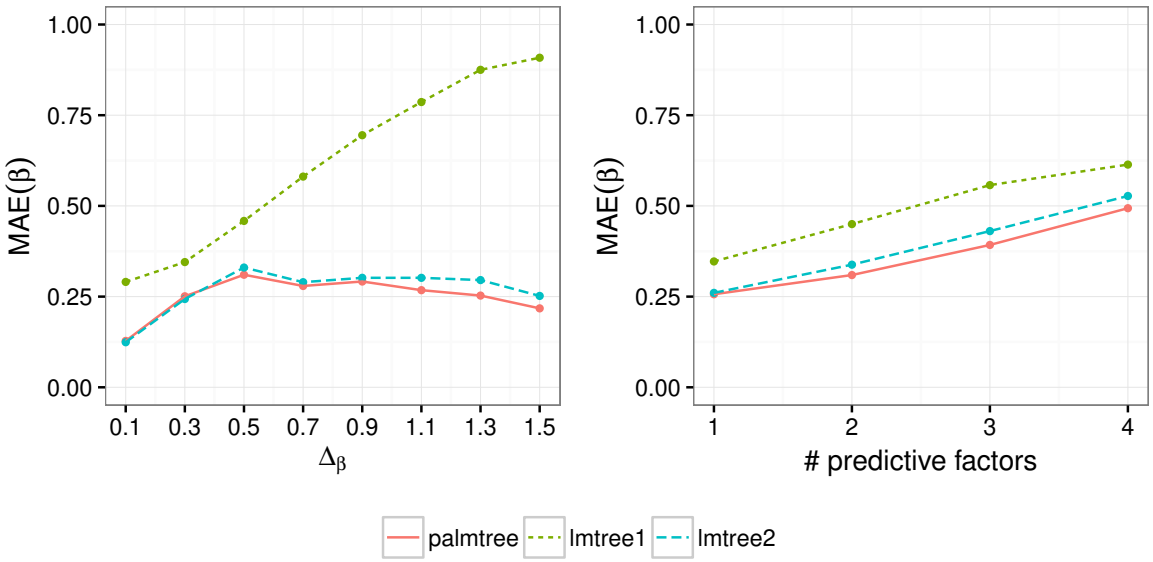


Figure 6: Mean absolute difference between true and estimated treatment effect (mean absolute error, MAE; Question 3.4).

random guessing. With increasing Δ_β all methods get better. The performance of PALM tree, LM tree 2 and OTR is similar, but OTR is not as good as expected. The three methods also behave similarly with a changing *number of predictive factors*. The treatment regime prediction is globally worst on average when there is one predictive factor. This results from the fact that often there is no split found (see Figure 3). When only one group is present, it automatically leads to a 50 percent chance of correct treatment assignment in the given simulation scenario. When there are two or three predictive factors, the proportion of patients for which the correct treatment is predicted to be the better treatment improves. For four predictive factors it is lower than for three for all methods.

3.4. How good is the treatment effect estimate?

Estimating or even predicting the correct treatment effect is the most essential part of subgroup analysis. Even if one treatment better is better than the other, clinicians need to know if the difference is relevant. The evaluation of the treatment effect estimate can only be done for the model-based recursive partitioning methods since OTR is only designed to produce binary decision rules. The measure used to evaluate the treatment effect estimate is the mean absolute difference between true and estimated treatment effect (mean absolute error, MAE). Figure 6 shows the MAE for the scenarios of varying Δ_β and varying number of predictive factors. The error is smallest for all three methods when the *difference in treatment effect* is lowest ($\Delta_\beta = 0.1$), because even if the chosen subgroups are wrong, the estimated treatment effect will likely be close to the true and very similar treatment effects. In this sense it is not a disadvantage that PALM tree and LM tree 2 often do not split into subgroups at all. In fact, it may even be an advantage, as the treatment effect estimate is then calculated based on a larger data set and is less affected by random variability. The effect of the small treatment difference gets less as the difference increases. However, as the difference increases, finding

the correct subgroups becomes easier and the error decreases. at the same time finding the correct subgroups becomes easier and slowly the error decreases again for PALM tree and LM tree 2. For this effect to be visible for LM tree 1, one would have to have even larger treatment effects, given the large effect of the prognostic factor. With an increasing *number of predictive factors* the mean absolute error in treatment effect increases. The shape of the curve looks very different than the one in Figure 5, even though they address similar questions, but the more true predictive factors exist in the given simulation scenario the harder it is for the methods to predict the treatment effect. This suggests that simply knowing the more effective treatment does not tell the whole story.

4. Illustration: Treatment differences in mathematics exam

The Mathematics 101 course for first-year business and economics students at Universität Innsbruck gives an introduction to mathematical analysis, linear algebra, financial mathematics, and probability calculus. Students are assessed by biweekly online tests during the semester and a written exam at the end. The exam consists of 13 single-choice questions with 5 answer alternatives, one of which is correct. Students who answer more than 60 percent of the questions correctly pass the course. The percentage of successful online tests captures math ability of the students and is a known predictor for success in the final exam.

The data contains the exam results of 729 students (out of 941 who originally registered for the course) for the fall semester in 2014/15. Due to limited availability of seats in the exam room, the students were asked to select a group, where the first group wrote the exam in the morning and the second group right after the first group finished. The two groups received slightly different questions on the same topics covering the scope of the course. We are interested in whether the exam is fair in the sense that it is on average equally hard or difficult for the two groups. In other words we want to find out whether there is a "treatment effect" with the different selection of exam questions in the two groups corresponding to the "treatments". As a first rather naive check we consider a simple one-way regression model for the percentage of correct answers by group, as reported in the first column of Table 2. This yields an expected percentage of 57.6 for a student in group 1 and a difference of 2.33 percentage points for students in group 2. Thus, the model finds only a small drop in the percentage of correctly solved answers and the corresponding confidence interval includes a zero change.

However, in this first model we have neglected the influence of the students' ability which is particularly relevant here because the students could freely choose their exam group. Therefore, there might have been self-selection of more (or less) able students into the first (or second) group. To account for such ability effects in the model we include the percentage of points from the previous online tests that captures the students' ability and preparation. As shown in the second column of Table 2 this variable is indeed strongly associated with the exam results, where one additional percentage point in the online tests leads to additional 0.86 expected percentage points in the written exam. More importantly, the group effect increases to 4.37 and the corresponding confidence interval does not include zero anymore. Despite the increase in the group effect, the absolute size of the group difference is still moderate corresponding to about half an exercise out of 13.

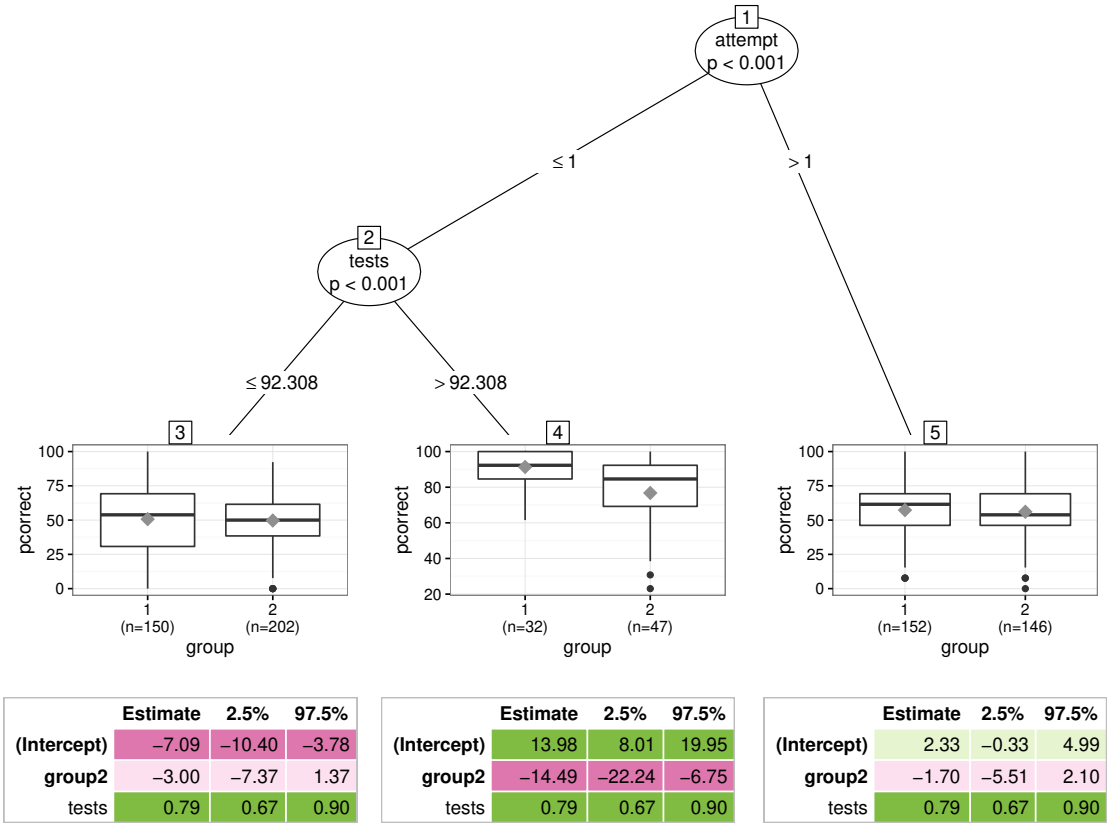
To explore the size of the treatment effect for the group differences further, we consider the

	Linear model 1	Linear model 2	PALM tree
(Intercept)	57.60 [55.12, 60.08]	−5.85 [−13.52, 1.83]	
node3:(Intercept)			−7.09 [−16.15, 1.97]
node4:(Intercept)			13.98 [0.82, 27.14]
node5:(Intercept)			2.33 [−6.32, 10.99]
group2	−2.33 [−5.70, 1.03]	−4.37 [−7.23, −1.50]	
node3:group2			−3.00 [−6.97, 0.98]
node4:group2			−14.49 [−22.92, −6.07]
node5:group2			−1.70 [−5.97, 2.56]
tests		0.86 [0.76, 0.95]	0.79 [0.67, 0.90]

Table 2: Three models for the mathematics exam data. The response variable is the percentage of correctly solved exercises and the main regressor of interest are the treatment differences between the first and second exam group. Confidence intervals are given in brackets.

possibility that this may vary across subgroups of students. Known student characteristics that may lead to such subgroups here are gender, the number of semesters the student has already been studying, the number of times the student has already attempted the exam, the type of study (three year bachelor program vs. four year diploma program) and also the ability/preparation as captured by percentage of successful exercises in the online tests. Figure 7 shows the resulting PALM tree with the segmented local group effect while adjusting for a global online tests effect. The strongest parameter instability is associated with the number of attempts and the group of students in the first attempt are split a second time by the percentage from the online tests. Two of the resulting subgroups (node 3 and 5) exhibit only very small group differences but in node 4 the second group obtained clearly a lower response percentage. This node is the smallest subgroup found and encompasses the highly able students taking the course for the first time. For this subsample the treatment effect is about 14 percentage points, which means that the students in the second batch solved about two exercises less than those in the first batch.

Overall this clearly conveys the strength of the PALM tree method: Especially in situations where the coefficient of interest is modest in a main-effects model and where further covariates are available whose influence on the main model parameters is not obvious, the PALM tree is an attractive option to globally control for certain variables while searching for local effects in others. Note, however, that due to the forward selection of models/effects the resulting confidence intervals in the terminal nodes (Table 2 and Figure 7) should not be used for inference but interpreted as a measure of variability.



	Estimate	2.5%	97.5%
(Intercept)	−7.09	−10.40	−3.78
group2	−3.00	−7.37	1.37
tests	0.79	0.67	0.90

	Estimate	2.5%	97.5%
(Intercept)	13.98	8.01	19.95
group2	−14.49	−22.24	−6.75
tests	0.79	0.67	0.90

	Estimate	2.5%	97.5%
(Intercept)	2.33	−0.33	4.99
group2	−1.70	−5.51	2.10
tests	0.79	0.67	0.90

Figure 7: PALM tree for the percentage of correct answers explained by group differences while globally adjusting for ability (i.e., percentage of points obtained in previous online tests).

5. Discussion

Model-based trees are effective tools to identify subgroups in data which differ in terms of model parameters. PALM trees are special model-based trees where some parameters can be fixed globally for the entire sample and do not depend on the subgroup structure. Our simulation study has shown that in cases where there are such specified factors with a direct effect on the outcome, PALM trees reliably detect the correct subgroups while at the same time having a low probability of detecting subgroups when there are none. Although optimal treatment regimes (OTR) perform comparably to PALM trees in terms of detecting the best treatment option in the given simulation study, they are typically better at recovering a parsimonious tree capturing the underlying subgroup structure. This makes PALM tree results easier to interpret and to communicate to practitioners, which we believe is an important advantage in many applications. Moreover, the simulation study clearly showed the effect of misspecifications in global vs. local effects in PALM trees. While it is important to correctly identify the variables with *additive* effects (LM tree 1 vs. LM tree 2 or PALM tree), it is not so important to correctly identify whether these additive effects are *global or local* (LM tree 2

vs. PALM tree). However, some power and efficiency can be gained from selecting a suitable PALM tree.

PALM trees allow exploring and questioning results of (generalised) linear models. The PALM tree analysis of the Mathematics 101 exam showed that a linear model regressing the percentage points of correct anwers on the group and earlier test results is too simple. Only for a relatively small subgroup of students who attempted the exam for the first time and who showed good performance during the semester it did make a difference whether they attempted the exam in the first or second group.

Although large parts of this manuscript focus on subgroup analyses in clinical trials, PALM trees can also be applied in a wide range of other applications as well – e.g., in the social sciences as shown in the mathematics exam application case study.

Computational details

Open-source implementations of the model-based tree algorithms LM tree, GLM tree and PALM tree are available in the **partykit** package (Hothorn and Zeileis 2015, functions `lmtree()`, `glmtree()` and `palmtree()`). The manuscript including simulation study and application can be reproduced using the material on `https://hub.docker.com/r/heidiseibold/palmtree-project/`.

Acknowledgements

We thank Andrea Farnham for improving the language. We are thankful to the Swiss National Fund for funding this project with grant 205321_163456 and mobility grant 205321_163456/2.

References

Breiman L, Friedman J, Stone CJ, Olshen RA (1984). *Classification and Regression Trees*. Wadsworth.

Chen J, Yu K, Hsing A, Therneau TM (2007). "A Partially Linear Tree-Based Regression Model for Assessing Complex Joint Gene-Gene and Gene-Environment Effects." *Genetic Epidemiology*, **31**(3), 238–251. doi:10.1002/gepi.20205.

Dusseldorp E, Conversano C, Van Os BJ (2010). "Combining an Additive and Tree-Based Regression Model Simultaneously: STIMA." *Journal of Computational and Graphical Statistics*, **19**(3), 514–530. doi:10.1198/jcgs.2010.06089.

Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H (2015). "Detecting Treatment-Subgroup Interactions in Clustered Data with Generalized Linear Mixed-Effects Model Trees." *Working Paper 2015-10*, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck. URL `http://EconPapers.RePEc.org/RePEc:inn:wpaper:2015-10`.

Hajjem A, Bellavance F, Larocque D (2011). "Mixed Effects Regression Trees for Clustered Data." *Statistics & Probability Letters*, **81**(4), 451–459. doi:10.1016/j.spl.2010.12.003.

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

Hubert L, Arabie P (1985). "Comparing Partitions." *Journal of Classification*, **2**(1), 193–218. doi:10.1007/BF01908075.

Italiano A (2011). "Prognostic or Predictive? It's Time to Get Back to Definitions!" *Journal of Clinical Oncology*, **29**(35), 4718–4718. doi:10.1200/JCO.2011.38.3729.

Lipkovich I, Dmitrienko A, D'Agostino RB (2016). "Tutorial in Biostatistics: Data-Driven Subgroup Identification and Analysis in Clinical Trials." *Statistics in Medicine*. doi:10.1002/sim.7064.

Mbogning C, Toussile W (2015). *GPLTR: Generalized Partially Linear Tree-Based Regression Model*. R package version 1.2, URL https://CRAN.R-project.org/package=GPLTR.

Milligan GW, Cooper MC (1986). "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis." *Multivariate Behavioral Research*, **21**(4), 441–458. doi:10.1207/s15327906mbr2104_5.

Sela RJ, Simonoff JS (2012). "RE-EM Trees: A Data Mining Approach for Longitudinal and Clustered Data." *Machine Learning*, **86**(2), 169–207. doi:10.1007/s10994-011-5258-3.

Sies A, Van Mechelen I (2016). "Comparing Four Methods for Estimating Tree-Based Treatment Regimes." Unpublished manuscript.

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514. doi:10.1198/106186008X319331.

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E (2012). "Estimating Optimal Treatment Regimes from a Classification Perspective." *Stat*, **1**(1), 103–114. doi:10.1002/sta.411.

Affiliation:

Heidi Seibold, Torsten Hothorn
Department of Biostatistics
Epidemiology, Biostatistics and Prevention Institute
University of Zurich
Hirschengraben 84
CH-8001 Zurich, Switzerland

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
A-6020 Innsbruck, Austria