Unbiased Recursive Partitioning: A Conditional Inference Framework

Torsten Hothorn
Friedrich-Alexander-Universität Erlangen-Nürnberg,

Kurt Hornik
Wirtschaftsuniversität Wien

Achim Zeileis
Wirtschaftsuniversität Wien

Abstract

Recursive binary partitioning is a popular tool for regression analysis. Two fundamental problems of exhaustive search procedures usually applied to fit such models have been known for a long time: overfitting and a selection bias towards covariates with many possible splits or missing values. While pruning procedures are able to solve the overfitting problem, the variable selection bias still seriously affects the interpretability of tree-structured regression models. For some special cases unbiased procedures have been suggested, however lacking a common theoretical foundation. We propose a unified framework for recursive partitioning which embeds tree-structured regression models into a well defined theory of conditional inference procedures. Stopping criteria based on multiple test procedures are implemented and it is shown that the predictive performance of the resulting trees is as good as the performance of established exhaustive search procedures. It turns out that the partitions and therefore the models induced by both approaches are structurally different, confirming the need for an unbiased variable selection. Moreover, it is shown that the prediction accuracy of trees with early stopping is equivalent to the prediction accuracy of pruned trees with unbiased variable selection. The methodology presented here is applicable to all kinds of regression problems, including nominal, ordinal, numeric, censored as well as multivariate response variables and arbitrary measurement scales of the covariates. Data from studies on glaucoma classification, node positive breast cancer survival and mammography experience are re-analyzed.

Keywords: permutation tests, variable selection, multiple testing, ordinal regression trees, multivariate regression trees.

1. Introduction

Statistical models that regress the distribution of a response variable on the status of multiple covariates are tools for handling two major problems in applied research: prediction and explanation. The function space represented by regression models focusing on the prediction problem may be arbitrarily complex; indeed, 'black box' systems like support vector machines or ensemble methods are excellent predictors. In contrast, regression models appropriate for gaining insight into the mechanism of the data generating process are required to offer a human readable representation. Generalized linear models or the Cox model are representatives of regression models where parameter estimates of the coefficients and their distribution are used to judge the relevance of single covariates.

With their seminal work on automated interaction detection (AID), Morgan and Sonquist (1963) introduced another class of simple regression models for prediction and explanation nowadays known as 'recursive partitioning' or 'trees'. Many variants and extensions have been published in the last 40 years, the majority of which are special cases of a simple two-stage algorithm: first partition the observations by univariate splits in a recursive way and second fit a constant model in each cell of the resulting partition. The most popular implementations of such algorithms are 'CART' (Breiman, Friedman, Olshen, and Stone 1984) and 'C4.5' (Quinlan 1993). Not unlike AID, both perform an exhaustive search over all possible splits maximizing an information measure of

This is a preprint of an article published in *Journal of Computational and Graphical Statistics*, Volume 15, Number 3, Pages 651–674.

node impurity selecting the covariate showing the best split. This approach has two fundamental problems: overfitting and a selection bias towards covariates with many possible splits. With respect to the overfitting problem Mingers (1987) notes that the algorithm

> [...] has no concept of statistical significance, and so cannot distinguish between a significant and an insignificant improvement in the information measure.

Within the exhaustive search framework, pruning procedures, mostly based on some form of cross-validation, are necessary to restrict the number of cells in the resulting partitions in order to avoid overfitting problems. While pruning is successful in selecting the right-sized tree, the interpretation of the trees is affected by the biased variable selection. This bias is induced by maximizing a splitting criterion over all possible splits simultaneously and was identified as a problem by many researchers (e.g., Kass 1980; Segal 1988; Breiman *et al.* 1984, p. 42). The nature of the variable selection problem under different circumstances has been studied intensively (White and Liu 1994; Jensen and Cohen 2000; Shih 2004) and Kim and Loh (2001) argue that exhaustive search methods are biased towards variables with many missing values as well. With this article we enter at the point where White and Liu (1994) demand for

> [...] a *statistical* approach [to recursive partitioning] which takes into account the *distributional* properties of the measures.

We present a unified framework embedding recursive binary partitioning with piecewise constant fits into the well-defined theory of permutation tests developed by Strasser and Weber (1999). The conditional distribution of statistics measuring the association between responses and covariates is the basis for an unbiased selection among covariates measured at different scales. Moreover, multiple test procedures are applied to determine whether no significant association between any of the covariates and the response can be stated and the recursion needs to stop. We show that such statistically motivated stopping criteria implemented via hypothesis tests lead to regression models whose predictive performance is equivalent to the performance of optimally pruned trees, therefore offering an intuitive and computationally efficient solution to the overfitting problem.

The development of the framework presented here was inspired by various attempts to solve both the overfitting and variable selection problem published in the last 25 years (a far more detailed overview is given by Murthy 1998). The χ^2 automated interaction detection algorithm ('CHAID', Kass 1980) is the first approach based on statistical significance tests for contingency tables. The basic idea of this algorithm is the separation of the variable selection and splitting procedure. The significance of the association between a nominal response and one of the covariates is investigated by a χ^2 test and the covariate with highest association is selected for splitting. Consequently, this algorithm has a concept of statistical significance and a criterion to stop the algorithm can easily be implemented based on formal hypothesis tests.

A series of papers aiming at unbiased recursive partitioning for nominal and continuous responses starts with 'FACT' (Loh and Vanichsetakul 1988), where covariates are selected within an analysis of variance (ANOVA) framework treating a nominal response as the independent variable. Basically, the covariate with largest F-ratio is selected for splitting. Nominal covariates are coerced to ordered variables via the canonical variate of the corresponding matrix of dummy codings. This induces a biased variable selection when nominal covariates are present and therefore 'QUEST' (Loh and Shih 1997) addresses this problem by selecting covariates on a P-value scale. For continuous variables, P-values are derived from the corresponding ANOVA F-statistics and for nominal covariates a χ^2 test is applied. This approach reduces the variable selection bias substantially. Further methodological developments within this framework include the incorporation of a linear discriminant analysis model within each node of a tree (Kim and Loh 2003) and multiway splits ('CRUISE', Kim and Loh 2001). For continuous responses, 'GUIDE' (Loh 2002) seeks to implement unbiasedness by a different approach. Here, the association between the sign of model residuals and each covariate is measured by a P-value derived from a χ^2 test. Continuous covariates are categorized to four levels prior to variable selection; however, models are fitted to untransformed

covariates in the nodes. These approaches are already very successful in reducing the variable selection bias and typically perform very well in the partitioning tasks they were designed for. Building on these ideas, we introduce a new unifying conceptual framework for unbiased recursive partitioning based on conditional hypothesis testing that, in addition to models for continuous and categorical data, includes procedures applicable to censored, ordinal or multivariate responses.

Previous attempts to implement permutation (or randomization) tests in recursive partitioning algorithms aimed at solving the variable selection and overfitting problem (Jensen and Cohen 2000), however focusing on special situations only. Resampling procedures have been employed for assessing split statistics for censored responses by LeBlanc and Crowley (1993). Frank and Witten (1998) utilize the conditional Monte-Carlo approach for the approximation of the distribution of Fisher's exact test for nominal responses and the conditional probability of an observed contingency table is used by Martin (1997). The asymptotic distribution of a 2×2 table obtained by maximizing the χ^2 statistic over possible splits in a continuous covariate is derived by Miller and Siegmund (1982). Maximally selected rank statistics (Lausen and Schumacher 1992) can be applied to continuous and censored responses as well and are applied to correct the bias of exhaustive search recursive partitioning by Lausen, Hothorn, Bretz, and Schumacher (2004). An approximation to the distribution of the Gini criterion is given by Dobra and Gehrke (2001). However, lacking solutions for more general situations, these auspicious approaches are hardly ever applied and the majority of tree-structured regression models reported and interpreted in applied research papers is biased. The main reason is that computationally efficient solutions are available for special cases only.

The framework presented in Section 3 is efficiently applicable to regression problems where both response and covariates can be measured at arbitrary scales, including nominal, ordinal, discrete and continuous as well as censored and multivariate variables. The treatment of special situations is explained in Section 4 and applications including glaucoma classification, node positive breast cancer survival and a questionnaire on mammography experience illustrate the methodology in Section 5. Finally, we show by benchmarking experiments that recursive partitioning based on statistical criteria as introduced in this paper lead to regression models whose predictive performance is as good as the performance of optimally pruned trees.

2. Recursive binary partitioning

We focus on regression models describing the conditional distribution of a response variable $\mathbf{Y}$ given the status of m covariates by means of tree-structured recursive partitioning. The response $\mathbf{Y}$ from some sample space $\mathcal{Y}$ may be multivariate as well. The m-dimensional covariate vector $\mathbf{X} = (X_1, \ldots, X_m)$ is taken from a sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$. Both response variable and covariates may be measured at arbitrary scales. We assume that the conditional distribution $D(\mathbf{Y}|\mathbf{X})$ of the response $\mathbf{Y}$ given the covariates $\mathbf{X}$ depends on a function f of the covariates

$$D(\mathbf{Y}|\mathbf{X}) = D(\mathbf{Y}|X_1, \ldots, X_m) = D(\mathbf{Y}|f(X_1, \ldots, X_m)),$$

where we restrict ourselves to partition based regression relationships, i.e., r disjoint cells $B_1, \ldots, B_r$ partitioning the covariate space $\mathcal{X} = \bigcup_{k=1}^{r} B_k$. A model of the regression relationship is to be fitted based on a learning sample $\mathcal{L}_n$, i.e., a random sample of n independent and identically distributed observations, possibly with some covariates X_{ji} missing,

$$\mathcal{L}_n \quad = \quad \{(\mathbf{Y}_i, X_{1i}, \ldots, X_{mi}); i = 1, \ldots, n\}.$$

A generic algorithm for recursive binary partitioning for a given learning sample $\mathcal{L}_n$ can be formulated using non-negative integer valued case weights $\mathbf{w} = (w_1, \ldots, w_n)$. Each node of a tree is represented by a vector of case weights having non-zero elements when the corresponding observations are elements of the node and are zero otherwise. The following generic algorithm implements recursive binary partitioning:

1. For case weights $\mathbf{w}$ test the global null hypothesis of independence between any of the m covariates and the response. Stop if this hypothesis cannot be rejected. Otherwise select the j^*th covariate X_{j^*} with strongest association to $\mathbf{Y}$.

2. Choose a set $A^* \subset \mathcal{X}_{j^*}$ in order to split $\mathcal{X}_{j^*}$ into two disjoint sets A^* and $\mathcal{X}_{j^*} \setminus A^*$. The case weights $\mathbf{w}_{\text{left}}$ and $\mathbf{w}_{\text{right}}$ determine the two subgroups with $w_{\text{left},i} = w_i I(X_{j^*i} \in A^*)$ and $w_{\text{right},i} = w_i I(X_{j^*i} \notin A^*)$ for all $i = 1, \ldots, n$ ($I(\cdot)$ denotes the indicator function).

3. Recursively repeat steps 1 and 2 with modified case weights $\mathbf{w}_{\text{left}}$ and $\mathbf{w}_{\text{right}}$, respectively.

As we sketched in the introduction, the separation of variable selection and splitting procedure into steps 1 and 2 of the algorithm is the key for the construction of interpretable tree structures not suffering a systematic tendency towards covariates with many possible splits or many missing values. In addition, a statistically motivated and intuitive stopping criterion can be implemented: We stop when the global null hypothesis of independence between the response and any of the m covariates cannot be rejected at a pre-specified nominal level α. The algorithm induces a partition $\{B_1, \ldots, B_r\}$ of the covariate space $\mathcal{X}$, where each cell $B \in \{B_1, \ldots, B_r\}$ is associated with a vector of case weights.

3. Recursive partitioning by conditional inference

In the main part of this section we focus on step 1 of the generic algorithm. Unified tests for independence are constructed by means of the conditional distribution of linear statistics in the permutation test framework developed by Strasser and Weber (1999). The determination of the best binary split in one selected covariate and the handling of missing values is performed based on standardized linear statistics within the same framework as well.

Variable selection and stopping criteria

At step 1 of the generic algorithm given in Section 2 we face an independence problem. We need to decide whether there is any information about the response variable covered by any of the m covariates. In each node identified by case weights $\mathbf{w}$, the global hypothesis of independence is formulated in terms of the m partial hypotheses $H_0^j : D(\mathbf{Y}|X_j) = D(\mathbf{Y})$ with global null hypothesis $H_0 = \bigcap_{j=1}^m H_0^j$. When we are not able to reject H_0 at a pre-specified level α, we stop the recursion. If the global hypothesis can be rejected, we measure the association between $\mathbf{Y}$ and each of the covariates $X_j, j = 1, \ldots, m$, by test statistics or P-values indicating the deviation from the partial hypotheses H_0^j.

For notational convenience and without loss of generality we assume that the case weights w_i are either zero or one. The symmetric group of all permutations of the elements of $(1, \ldots, n)$ with corresponding case weights $w_i = 1$ is denoted by $S(\mathcal{L}_n, \mathbf{w})$. A more general notation is given in Appendix A. We measure the association between $\mathbf{Y}$ and $X_j, j = 1, \ldots, m$, by linear statistics of the form

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = \text{vec}\left(\sum_{i=1}^n w_i g_j(X_{ji}) h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n))^\top\right) \in \mathbb{R}^{p_j q} \tag{1}$$

where $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ is a non-random transformation of the covariate X_j. The *influence function* $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ depends on the responses $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ in a permutation symmetric way. Section 4 explains how to choose g_j and h in different practical settings. A $p_j \times q$ matrix is converted into a $p_j q$ column vector by column-wise combination using the 'vec' operator.

The distribution of $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ under H_0^j depends on the joint distribution of $\mathbf{Y}$ and X_j, which is unknown under almost all practical circumstances. At least under the null hypothesis one can dispose of this dependency by fixing the covariates and conditioning on all possible permutations of the responses. This principle leads to test procedures known as *permutation tests*. The conditional

expectation $\mu_j \in \mathbb{R}^{p_j q}$ and covariance $\Sigma_j \in \mathbb{R}^{p_j q \times p_j q}$ of $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ under H_0 given all permutations $\sigma \in S(\mathcal{L}_n, \mathbf{w})$ of the responses are derived by Strasser and Weber (1999):

$$\begin{aligned}
\mu_j &= \mathbb{E}(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})|S(\mathcal{L}_n, \mathbf{w})) = \operatorname{vec}\left(\left(\sum_{i=1}^n w_i g_j(X_{ji})\right)\mathbb{E}(h|S(\mathcal{L}_n, \mathbf{w}))^\top\right), \\
\Sigma_j &= \mathbb{V}(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})|S(\mathcal{L}_n, \mathbf{w})) \\
&= \frac{\mathbf{w}_\cdot}{\mathbf{w}_\cdot - 1}\mathbb{V}(h|S(\mathcal{L}_n, \mathbf{w})) \otimes \left(\sum_i w_i g_j(X_{ji}) \otimes w_i g_j(X_{ji})^\top\right) \\
&- \frac{1}{\mathbf{w}_\cdot - 1}\mathbb{V}(h|S(\mathcal{L}_n, \mathbf{w})) \otimes \left(\sum_i w_i g_j(X_{ji})\right) \otimes \left(\sum_i w_i g_j(X_{ji})\right)^\top
\end{aligned} \tag{2}$$

where $\mathbf{w}_\cdot = \sum_{i=1}^n w_i$ denotes the sum of the case weights, $\otimes$ is the Kronecker product and the conditional expectation of the influence function is

$$\mathbb{E}(h|S(\mathcal{L}_n, \mathbf{w})) = \mathbf{w}_\cdot^{-1} \sum_i w_i h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) \in \mathbb{R}^q$$

with corresponding $q \times q$ covariance matrix

$$\begin{aligned}
\mathbb{V}(h|S(\mathcal{L}_n, \mathbf{w})) = \mathbf{w}_\cdot^{-1} \sum_i w_i &\left(h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) - \mathbb{E}(h|S(\mathcal{L}_n, \mathbf{w}))\right) \\
&\left(h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) - \mathbb{E}(h|S(\mathcal{L}_n, \mathbf{w}))\right)^\top.
\end{aligned}$$

Having the conditional expectation and covariance at hand we are able to standardize a linear statistic $\mathbf{T} \in \mathbb{R}^{pq}$ of the form (1) for some $p \in \{p_1, \ldots, p_m\}$. Univariate test statistics c mapping an observed multivariate linear statistic $\mathbf{t} \in \mathbb{R}^{pq}$ into the real line can be of arbitrary form. An obvious choice is the maximum of the absolute values of the standardized linear statistic

$$c_{\max}(\mathbf{t}, \mu, \Sigma) = \max_{k=1,\ldots,pq} \left|\frac{(\mathbf{t} - \mu)_k}{\sqrt{(\Sigma)_{kk}}}\right|$$

utilizing the conditional expectation μ and covariance matrix Σ. The application of a quadratic form $c_{\text{quad}}(\mathbf{t}, \mu, \Sigma) = (\mathbf{t}-\mu)\Sigma^+(\mathbf{t}-\mu)^\top$ is one alternative, although computationally more expensive because the Moore-Penrose inverse Σ^+ of Σ is involved. It is important to note that the test statistics $c(\mathbf{t}_j, \mu_j, \Sigma_j), j = 1, \ldots, m$, cannot be directly compared in an unbiased way unless all of the covariates are measured at the same scale, i.e., $p_1 = p_j, j = 2, \ldots, m$. In order to allow for an unbiased variable selection we need to switch to the P-value scale because P-values for the conditional distribution of test statistics $c(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}), \mu_j, \Sigma_j)$ can be directly compared among covariates measured at different scales. In step 1 of the generic algorithm we select the covariate with minimum P-value, i.e., the covariate X_{j^*} with $j^* = \operatorname{argmin}_{j=1,\ldots,m} P_j$, where

$$P_j = \mathbb{P}_{H_0^j}(c(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}), \mu_j, \Sigma_j) \geq c(\mathbf{t}_j, \mu_j, \Sigma_j)|S(\mathcal{L}_n, \mathbf{w}))$$

denotes the P-value of the conditional test for H_0^j.

So far, we have only addressed testing each partial hypothesis H_0^j, which is sufficient for an unbiased variable selection. A global test for H_0 required in step 1 can be constructed via an aggregation of the transformations $g_j, j = 1, \ldots, m$, i.e., using a linear statistic of the form

$$\mathbf{T}(\mathcal{L}_n, \mathbf{w}) = \operatorname{vec}\left(\sum_{i=1}^n w_i \left(g_1(X_{1i})^\top, \ldots, g_m(X_{mi})^\top\right)^\top h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n))^\top\right).$$

However, this approach is less attractive for learning samples with missing values. Universally applicable approaches are multiple test procedures based on $P_1, \ldots, P_m$. Simple Bonferroni-adjusted P-values or a min-P-value resampling approach are just examples and we refer to the multiple testing literature (e.g., Westfall and Young 1993) for more advanced methods. We reject H_0 when the minimum of the adjusted P-values is less than a pre-specified nominal level α and otherwise stop the algorithm. In this sense, α may be seen as a unique parameter determining the size of the resulting trees.

The conditional distribution and thus the P-value of the statistic $c(\mathbf{t}, \mu, \Sigma)$ can be computed in several different ways (see Hothorn, Hornik, van de Wiel, and Zeileis 2006, for an overview). For some special forms of the linear statistic, the exact distribution of the test statistic is tractable; conditional Monte-Carlo procedures can always be used to approximate the exact distribution. Strasser and Weber (1999) proved (Theorem 2.3) that the conditional distribution of linear statistics $\mathbf{T}$ with conditional expectation μ and covariance Σ tends to a multivariate normal distribution with parameters μ and Σ as $n, \mathbf{w}. \to \infty$. Thus, the asymptotic conditional distribution of test statistics of the form $c_{\max}$ is normal and can be computed directly in the univariate case ($p_j q = 1$) or approximated by means of quasi-randomized Monte-Carlo procedures in the multivariate setting (Genz 1992). Quadratic forms c_{quad} follow a asymptotic χ^2 distribution with degrees of freedom given by the rank of Σ (Theorem 6.20, Rasch 1995), and therefore asymptotic P-values can be computed efficiently.

Splitting criteria

Once we have selected a covariate in step 1 of the algorithm, the split itself can be established by any splitting criterion, including those established by Breiman *et al.* (1984) or Shih (1999). Instead of simple binary splits, multiway splits can be implemented as well, for example utilizing the work of O'Brien (2004). However, most splitting criteria are not applicable to response variables measured at arbitrary scales and we therefore utilize the permutation test framework described above to find the optimal binary split in one selected covariate X_{j^*} in step 2 of the generic algorithm. The goodness of a split is evaluated by two-sample linear statistics which are special cases of the linear statistic (1). For all possible subsets A of the sample space $\mathcal{X}_{j^*}$ the linear statistic

$$\mathbf{T}^A_{j^*}(\mathcal{L}_n, \mathbf{w}) = \text{vec}\left(\sum_{i=1}^{n} w_i I(X_{j^* i} \in A) h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n))^\top\right) \in \mathbb{R}^q$$

induces a two-sample statistic measuring the discrepancy between the samples $\{\mathbf{Y}_i | w_i > 0 \text{ and } X_{ji} \in A; i = 1, \ldots, n\}$ and $\{\mathbf{Y}_i | w_i > 0 \text{ and } X_{ji} \notin A; i = 1, \ldots, n\}$. The conditional expectation $\mu^A_{j^*}$ and covariance $\Sigma^A_{j^*}$ can be computed by (2). The split A^* with a test statistic maximized over all possible subsets A is established:

$$A^* = \underset{A}{\text{argmax}}\, c(\mathbf{t}^A_{j^*}, \mu^A_{j^*}, \Sigma^A_{j^*}). \tag{3}$$

Note that we do not need to compute the distribution of $c(\mathbf{t}^A_{j^*}, \mu^A_{j^*}, \Sigma^A_{j^*})$ in step 2. In order to prevent pathological splits one can restrict the number of possible subsets that are evaluated, for example by introducing restrictions on the sample size or the sum of the case weights in each of the two groups of observations induced by a possible split.

Missing values and surrogate splits

If an observation X_{ji} in covariate X_j is missing, we set the corresponding case weight w_i to zero for the computation of $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ and, if we would like to split in X_j, in $\mathbf{T}^A_j(\mathcal{L}_n, \mathbf{w})$ as well. Once a split A^* in X_j has been implemented, surrogate splits can be established by searching for a split leading to roughly the same division of the observations as the original split. One simply replaces the original response variable by a binary variable $I(X_{ji} \in A^*)$ coding the split and proceeds as described in the previous part.

Choice of α

The parameter α can be interpreted in two different ways: as pre-specified nominal level of the underlying association tests or as a simple hyper parameter determining the tree size. In the first sense, α controls the probability of falsely rejecting H_0 in each node. The typical conventions for balancing the type I and type II errors apply in this situation.

Although the test procedures used for constructing the tree are general independence tests, they will only have high power for very specific directions of deviation from independence (depending on the choice of g and h) and lower power for any other direction of departure. Hence, a strategy to assure that any type of dependence is detected could be to increase the significance level α. To avoid that the tree grown with a very large α overfits the data, a final step could be added for pruning the tree in a variety of ways, for example by eliminating all terminal nodes until the terminal splits are significant at level α', with α' being much smaller than the initial α. Note, that by doing so the interpretation of α as nominal significance level of conditional test procedures is lost. Moreover, α can be seen as a hyper parameter that is subject to optimization with respect to some risk estimate, e.g., computed via cross-validation or additional test samples.

For explanatory modelling, the view of α as a significance level seems more intuitive and easier to explain to subject matter scientists, whereas for predictive modelling the view of α as a hyper parameter is also feasible. Throughout the paper we adopt the first approach and also evaluate it in a predictive setting in Section 6.

Computational complexity

The computational complexity of the variable selection step is of order n (for fixed $p_j, j = 1, \ldots, m$ and q) since computing each $\mathbf{T}_j$ with corresponding μ_j and Σ_j can be performed in linear time. The computations of the test statistics c is independent of the number of observations. Searching the optimal splits in continuous variables involves ranking these and hence is of order $n \log n$. However, for nominal covariates measured at K levels, the evaluation of all $2^{K-1} - 1$ possible splits is not necessary for the variable selection.

4. Examples

Univariate continuous or discrete regression

For a univariate numeric response $\mathbf{Y} \in \mathbb{R}$, the most natural influence function is the identity $h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) = \mathbf{Y}_i$. In cases where some observations with extremely large or small values have been observed, a ranking of the observations may be appropriate: $h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) = \sum_{k=1}^{n} w_k I(\mathbf{Y}_k \leq \mathbf{Y}_i)$ for $i = 1, \ldots, n$. Numeric covariates can be handled by the identity transformation $g_{ji}(x) = x$ (ranks or non-linear transformations are possible, too). Nominal covariates at levels $1, \ldots, K$ are represented by $g_{ji}(k) = e_K(k)$, the unit vector of length K with kth element being equal to one. Due to this flexibility, special test procedures like the Spearman test, the Wilcoxon-Mann-Whitney test or the Kruskal-Wallis test and permutation tests based on ANOVA statistics or correlation coefficients are covered by this framework. Splits obtained from (3) maximize the absolute value of the standardized difference between two means of the values of the influence functions. For prediction, one is usually interested in an estimate of the expectation of the response $\mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ in each cell; an estimate can be obtained by

$$\hat{\mathbb{E}}(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \left(\sum_{i=1}^{n} w_i(\mathbf{x})\right)^{-1} \sum_{i=1}^{n} w_i(\mathbf{x}) \mathbf{Y}_i,$$

where $w_i(\mathbf{x}) = w_i$ when $\mathbf{x}$ is element of the same terminal node as the ith observation and zero otherwise.

Censored regression

The influence function h may be chosen as logrank or Savage scores taking censoring into account and one can proceed as for univariate continuous regression. This is essentially the approach first published by Segal (1988). An alternative is the weighting scheme suggested by Molinaro, Dudoit, and van der Laan (2004). A weighted Kaplan-Meier curve for the case weights $\mathbf{w}(\mathbf{x})$ can serve as prediction.

J-Class classification

The nominal response variable at levels $1, \dots, J$ is handled by influence functions $h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) = e_J(\mathbf{Y}_i)$. Note that for a nominal covariate X_j at levels $1, \dots, K$ with $g_{ji}(k) = e_K(k)$ the corresponding linear statistic $\mathbf{T}_j$ is a vectorized contingency table of X_j and $\mathbf{Y}$. The conditional class probabilities can be estimated via

$$\hat{\mathbb{P}}(\mathbf{Y} = y | \mathbf{X} = \mathbf{x}) = \left(\sum_{i=1}^{n} w_i(\mathbf{x}) \right)^{-1} \sum_{i=1}^{n} w_i(\mathbf{x}) I(\mathbf{Y}_i = y), \quad y = 1, \dots, J.$$

Ordinal regression

Ordinal response variables measured at J levels, and ordinal covariates measured at K levels, are associated with score vectors $\xi \in \mathbb{R}^J$ and $\gamma \in \mathbb{R}^K$, respectively. Those scores reflect the 'distances' between the levels: If the variable is derived from an underlying continuous variable, the scores can be chosen as the midpoints of the intervals defining the levels. The linear statistic is now a linear combination of the linear statistic $\mathbf{T}_j$ of the form

$$\mathbf{MT}_j(\mathcal{L}_n, \mathbf{w}) = \operatorname{vec}\left(\sum_{i=1}^{n} w_i \gamma^\top g_j(X_{ji}) \left(\xi^\top h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) \right)^\top \right)$$

with $g_j(x) = e_K(x)$ and $h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) = e_J(\mathbf{Y}_i)$. If both response and covariate are ordinal, the matrix of coefficients is given by the Kronecker product of both score vectors $\mathbf{M} = \xi \otimes \gamma \in \mathbb{R}^{1, KJ}$. In case the response is ordinal only, the matrix of coefficients $\mathbf{M}$ is a block matrix

$$\mathbf{M} = \left(\begin{array}{ccc|c|ccc} \xi_1 & & 0 & & \xi_q & & 0 \\ & \ddots & & \dots & & \ddots & \\ 0 & & \xi_1 & & 0 & & \xi_q \end{array} \right) \text{ or } \mathbf{M} = \operatorname{diag}(\gamma)$$

when one covariate is ordered but the response is not. For both $\mathbf{Y}$ and X_j being ordinal, the corresponding test is known as linear-by-linear association test (Agresti 2002).

Multivariate regression

For multivariate responses, the influence function is a combination of influence functions appropriate for any of the univariate response variables discussed in the previous paragraphs, e.g., indicators for multiple binary responses (Zhang 1998; Noh, Song, and Park 2004), logrank or Savage scores for multiple failure times and the original observations or a rank transformation for multivariate regression (De'ath 2002).

5. Illustrations and applications

In this section, we present regression problems which illustrate the potential fields of application of the methodology. Conditional inference trees based on c_{quad}-type test statistics using the identity influence function for numeric responses and asymptotic χ^2 distribution are applied. For the

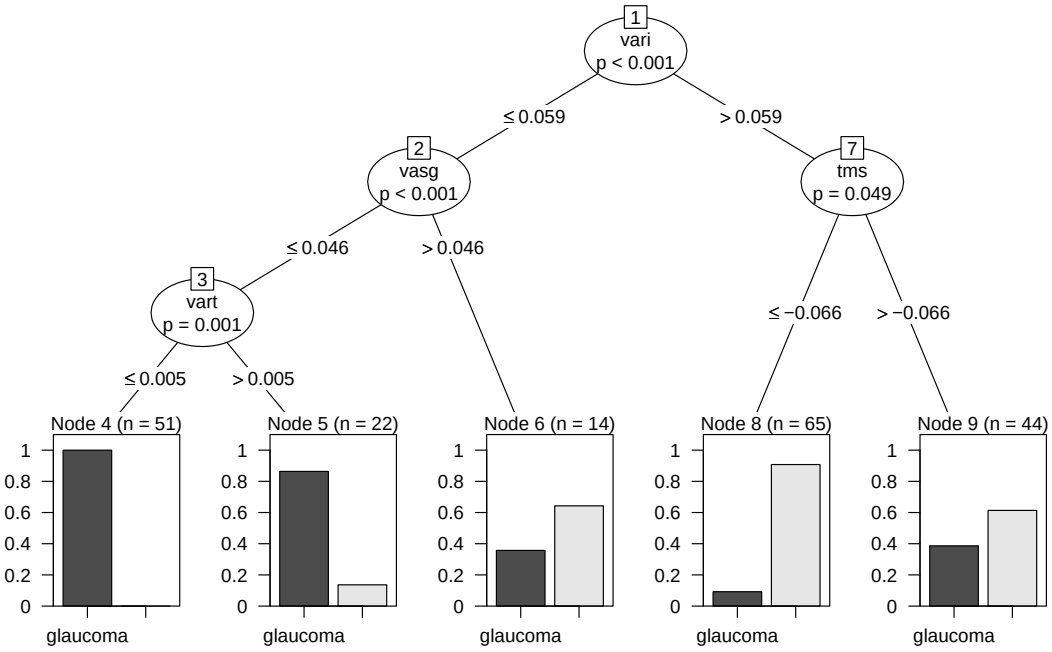

Figure 1: Conditional inference tree for the glaucoma data. For each inner node, the Bonferroni-adjusted P-values are given, the fraction of glaucomateous eyes is displayed for each terminal node.

stopping criterion a simple Bonferroni correction is used and we follow the usual convention by choosing the nominal level of the conditional independence tests as $\alpha = 0.05$. Conditional inference trees are implemented in the **party** add-on package to the R system for statistical computing (version 2.0.1, R Development Core Team 2004), both being freely available from CRAN (`http://CRAN.R-project.org/`). Our analyses can be reproduced using the code given in Appendix B.

Glaucoma and laser scanning images

Laser scanning images taken from the eye background are expected to serve as the basis of an automated system for glaucoma diagnosis. Although prediction is more important in this application (Mardin, Hothorn, Peters, Jünemann, Nguyen, and Lausen 2003), a simple visualization of the regression relationship is useful for comparing the structures inherent in the learning sample with subject matter knowledge. For 98 patients and 98 controls, matched by age and gender, 62 covariates describing the eye morphology are available. The data is part of the **ipred** package (Peters, Hothorn, and Lausen 2002). The first split in Figure 1 separates eyes with a volume above reference less than 0.059 mm^3 in the inferior part of the optic nerve head (`vari`). Observations with larger volume are mostly controls, a finding which corresponds to subject matter knowledge: The volume above reference measures the thickness of the nerve layer, expected to decrease with a glaucomateous damage of the optic nerve. Further separation is achieved by the volume above surface global (`vasg`) and the volume above reference in the temporal part of the optic nerve head (`vart`).

Node positive breast cancer

Recursive partitioning for censored responses has attracted a lot of interest (e.g., Segal 1988; LeBlanc and Crowley 1992). Survival trees using P-value adjusted logrank statistics are used by Schumacher, Holländer, Schwarzer, and Sauerbrei (2001) for the evaluation of prognostic factors

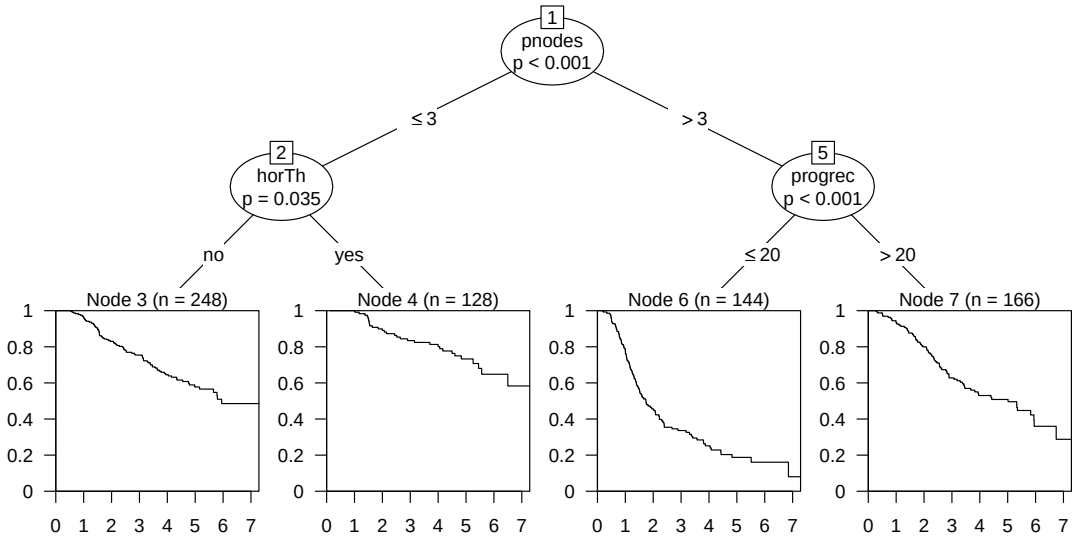

Figure 2: Tree-structured survival model for the GBSG2 data with Kaplan-Meier estimates of the survival time (in years) in the terminal nodes.

for the German Breast Cancer Study Group (GBSG2) data, a prospective controlled clinical trial on the treatment of node positive breast cancer patients. Here, we use logrank scores as well. Complete data of seven prognostic factors of 686 women are used for prognostic modeling, the dataset is available within the **ipred** package. The number of positive lymph nodes (`pnodes`) and the progesterone receptor (`progrec`) have been identified as prognostic factors in the survival tree analysis by Schumacher *et al.* (2001). Here, the binary variable coding whether a hormonal therapy was applied or not (`horTh`) additionally is part of the model depicted in Figure 2.

Mammography experience

Ordinal response variables are common in investigations where the response is a subjective human interpretation. We use an example given by Hosmer and Lemeshow (2000), p. 264, studying the relationship between the mammography experience (never, within a year, over one year) and opinions about mammography expressed in questionnaires answered by $n = 412$ women. The resulting partition based on scores $\xi = (1, 2, 3)$ is given in Figure 3. Most women who (strongly) agree with the question 'You do not need a mammogram unless you develop symptoms' have not experienced a mammography. The variable `benefit` is a score with low values indicating a strong agreement with the benefits of the examination. For those women in (strong) disagreement with the first question above, low values of `benefit` identify persons being more likely to have experienced such an examination at all.

6. Empirical comparisons

In this section, we investigate both the estimation and prediction accuracy of the conditional inference trees suggested in this paper. Three assertions are to be tested by means of benchmark experiments: 1) conditional inference trees are unbiased, 2) conditional inference trees do not suffer from overfitting and 3) the prediction accuracy of conditional inference trees is equivalent to the prediction accuracy of optimally pruned trees.

The **rpart**, **QUEST** and **GUIDE** software implementations serve as competitors for the comparisons. The **rpart** package (Therneau and Atkinson 1997) essentially implements the algorithms

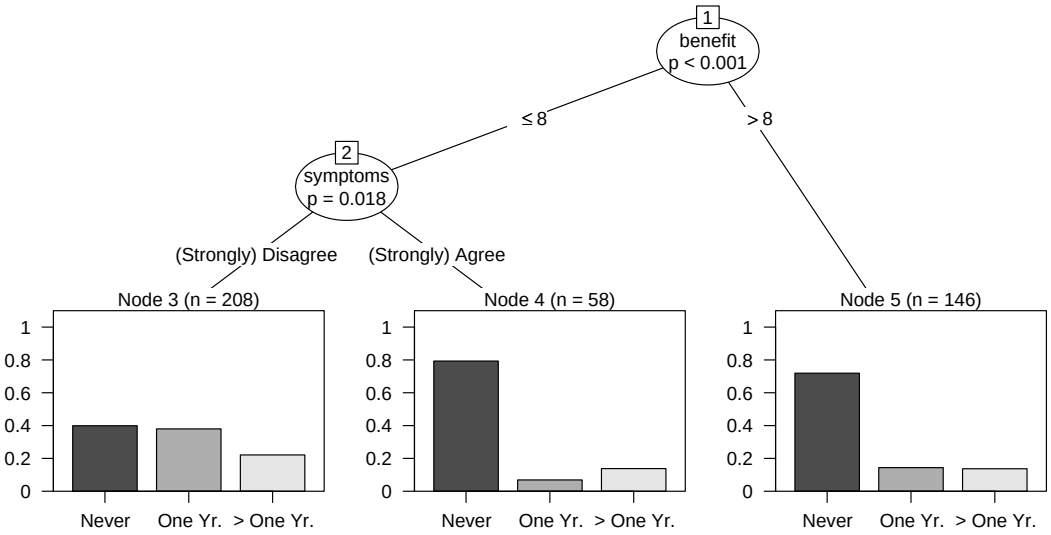

Figure 3: Ordinal regression for the mammography experience data with the fractions of (never, within a year, over one year) given in the nodes.

described in the CART book by Breiman *et al.* (1984) and is the de-facto standard in open-source recursive partitioning software. It implements cost-complexity pruning based on cross-validation after an initial large tree was grown by exhaustive search. **QUEST** (quick, unbiased and efficient statistical tree for nominal responses, Loh and Shih 1997), version 1.9.1, and **GUIDE** (generalized, unbiased, interaction detection and estimation for numeric responses, Loh 2002), version 2.1, aim at unbiased variable selection and determine the tree size by pruning as well. For the comparisons between conditional inference trees and **GUIDE**, the latter is limited to fitting constant models within each terminal node such that all algorithms fit a model from the same model class. We use binaries of both implementations available from `http://www.stat.wisc.edu/~loh/`.

The conditional inference trees are constructed with c_{quad}-type test statistics and $\alpha = 0.05$ with simple Bonferroni correction. Each split needs to send at least 1% of the observations into each of the two daughter nodes. The sample size in each node is restricted to 20 observations for all four algorithms under test, otherwise, the default settings of **rpart**, **QUEST** and **GUIDE** were not changed.

Estimation accuracy

The assertions 1) and 2) are tested by means of a simple simulation experiment, following the approach of Kim and Loh (2001) who demonstrate the unbiasedness of CRUISE empirically. An algorithm for recursive partitioning is called unbiased when, under the conditions of the null hypothesis of independence between a response $\mathbf{Y}$ and covariates $X_1, \ldots, X_m$, the probability of selecting covariate X_j is $1/m$ for all $j = 1, \ldots, m$ regardless of the measurement scales or number of missing values.

Five uniformly distributed random variables $X_1, \ldots, X_5 \sim \mathcal{U}[0, 1]$ serve as numeric covariates. In covariate X_4, 25% of the values are drawn missing at random, and the values of covariate X_5 are rounded to one digit, i.e., we induce 11 unique realizations. An additional nominal covariate X_6 is measured at two levels, with 50% of the observations being equal to zero. In this simple regression problem, the response variable $\mathbf{Y}$ is normal with means zero and μ in the two groups defined by covariate X_6.

	rpart		Conditional Inference Trees	
	Estimate	95% Confidence Interval	Estimate	95% Confidence Interval
$X_1 \sim \mathcal{U}[0,1]$	0.231	(0.220, 0.243)	0.168	(0.159, 0.178)
$X_2 \sim \mathcal{U}[0,1]$	0.225	(0.214, 0.236)	0.167	(0.157, 0.177)
$X_3 \sim \mathcal{U}[0,1]$	0.227	(0.216, 0.238)	0.162	(0.153, 0.172)
X_4, missings	0.197	(0.187, 0.208)	0.169	(0.159, 0.179)
X_5, ties	0.100	(0.092, 0.108)	0.166	(0.156, 0.176)
X_6, binary	0.020	(0.017, 0.024)	0.169	(0.159, 0.179)

Table 1: Simulated probabilities of variable selection of six mutually independent variables when the response is independent of $X_1, \ldots, X_6$, i.e., $\mu = 0$. The results are based on 10,000 replications.

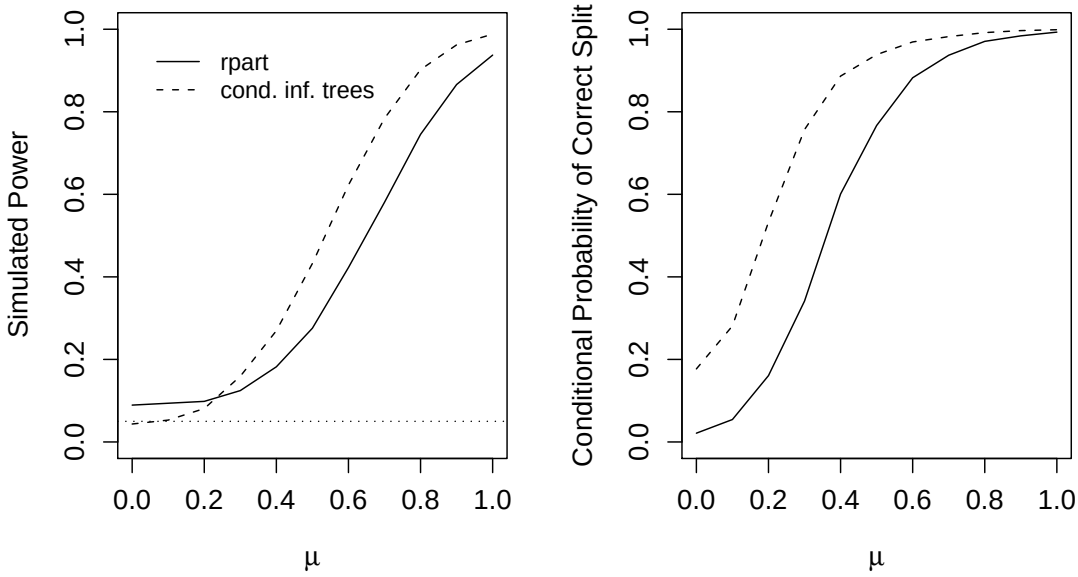

Figure 4: Simulated power, i.e. the probability of a root split (left), and the simulated conditional probability of a correct split in variable X_6 given that any root split was established (right) are displayed. The dotted horizontal line represents $\alpha = 0.05$. The results are based on 10,000 replications.

$$\mathbf{Y} \sim \begin{cases} \mathcal{N}(0,1) & \text{if } X_6 = 0 \\ \mathcal{N}(\mu,1) & \text{if } X_6 = 1. \end{cases}$$

For $\mu = 0$, the response is independent of all covariates. The probability of selecting $X_j, j = 1, \ldots, 6$, based on learning samples of size $n = 100$ drawn from the model above is estimated for both **rpart** and conditional inference trees by means of 10,000 simulation runs. Note that the root split is forced, i.e., no stopping criterion is applied for this experiment. The estimated probabilities in Table 1 illustrate the well-known fact that exhaustive search procedures, like **rpart**, are heavily biased towards covariates with many possible splits. The 95% simultaneous confidence intervals for the proportions (as described by Goodman 1965) for **rpart** never include 1/6. In contrast, the

confidence intervals for the conditional inference trees always include the probability 1/6 expected for an unbiased variable selection, regardless of the measurement scale of the covariates. This result indicates that the selection of covariates by asymptotic P-values of conditional independence tests is unbiased.

From a practical point of view, two issues with greater relevance arise. On the one hand, the probability of selecting any of the covariates for splitting for some $\mu \geq 0$ (power) and, on the other hand, the conditional probability of selecting the "correct split" in covariate X_6 given any covariate was selected for splitting are interesting criteria with respect to which the two algorithms are compared. Figure 4 depicts the estimated probabilities for varying μ. For $\mu = 0$, the probability of splitting the root node is 0.0435 for conditional inference trees and 0.0893 for **rpart**. Thus, the probability of such an incorrect decision is bounded by α for the conditional inference trees and is twice as large for pruning as implemented in **rpart**. Under the alternative $\mu > 0$, the conditional inference trees are more powerful compared to **rpart** for $\mu > 0.2$. For small values of μ the larger power of **rpart** is due to the size distortion under the null hypothesis. In addition, the probability of selecting X_6 given that any covariate was selected is uniformly greater for the conditional inference trees.

The advantageous properties of the conditional inference trees are obvious for the simple simulation model with one split only. We now extend our investigations to a simple regression tree with four terminal nodes. The response variable is normal with mean μ depending on the covariates as follows:

$$\mathbf{Y} \sim \begin{cases} \mathcal{N}(1,1) & \text{if } X_6 = 0 \text{ and } X_1 < 0.5 \\ \mathcal{N}(2,1) & \text{if } X_6 = 0 \text{ and } X_1 \geq 0.5 \\ \mathcal{N}(3,1) & \text{if } X_6 = 1 \text{ and } X_2 < 0.5 \\ \mathcal{N}(4,1) & \text{if } X_6 = 1 \text{ and } X_2 \geq 0.5. \end{cases} \tag{4}$$

We will focus on two closely related criteria describing the partitions induced by the algorithms: the complexity of the induced partitions and the structure of the trees. The number of terminal nodes of a tree is a measure of the complexity of the model and can easily be compared with the number of cells in the true data partition defined by (4). However, the appropriate complexity of a tree does not ensure that the tree structure describes the true data partition well. Here, we measure the discrepancy between the true data partition and the partitions obtained from recursive partitioning by the normalized mutual information ('NMI', Strehl and Ghosh 2003), essentially the mutual information of two partitions standardized by the entropy of both partitions. Values near one indicate similar to equal partitions while values near zero are obtained for structurally different partitions.

For 1,000 learning samples of size $n = 100$ drawn from the simple tree model, Table 2 gives the cross-tabulated number of terminal nodes of conditional inference trees and pruned exhaustive search trees computed by **rpart**. The null hypothesis of marginal homogeneity for ordered variables

		Conditional Inference Trees						
		2	3	4	5	6	≥ 7	
rpart	2	3	4	5	0	0	0	12
	3	0	48	47	3	0	0	98
	4	0	36	549	49	3	0	637
	5	0	12	134	25	1	0	172
	6	2	6	42	10	1	0	61
	≥ 7	0	3	10	6	1	0	20
		5	109	787	93	6	0	1000

Table 2: Number of terminal nodes for **rpart** and conditional inference trees when the learning sample is actually partitioned into four cells.

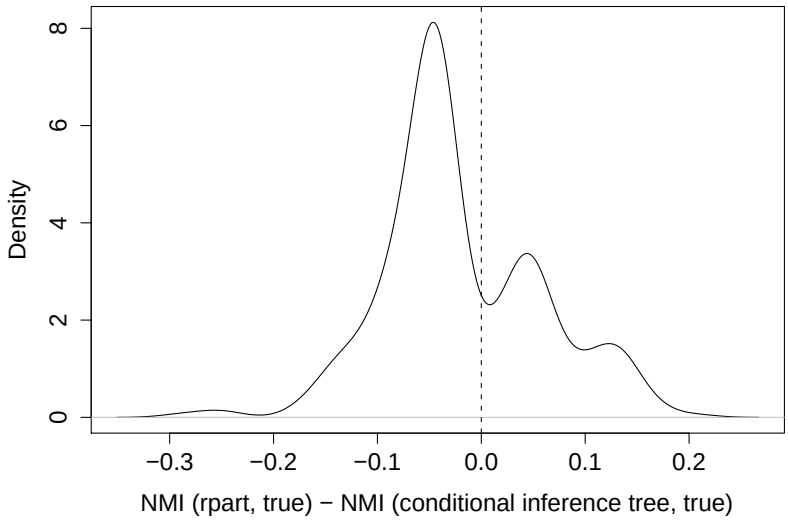

Figure 5: Density estimate of the difference in normalized mutual information of the true partition and the partitions induced by **rpart** and conditional inference trees. Instances with a NMI difference of zero were excluded – the results are based on 394 replications.

(Agresti 2002) can be rejected (P-value < 0.0001) indicating that the partitions obtained from both algorithms differ with respect to the number of terminal nodes. Conditional inference trees select a right-sized tree (four terminal nodes) in 78.7% of the cases while **rpart** generates trees with four terminal nodes for 63.7% of the learning samples. In general, pruning as implemented in **rpart** tends to produce trees with a larger number of terminal nodes in this example.

The correct tree structure with four leaves, with the first split in X_6 and splits in X_1 and X_2 in the left or right node, is detected by **rpart** in 63.3% of the simulation runs and in 77.5% of the cases by conditional inference trees. The NMI measure between the true partition of the data given by (4) and the partitions induced by the tree algorithms needs to be compared for instances with informative NMI measures only, i.e., the cases where the NMI between **rpart** and the true data partition and the NMI between conditional inference trees and the true data partition coincide do not cover any information. A density estimate of the NMI difference between partitions obtained from **rpart** and conditional inference tree partitions in Figure 5 shows that the partitions induced by conditional inference trees are, one average, closer to the true data partition.

Prediction accuracy

Assertion 3) is investigated by means of 11 benchmarking problems from the UCI repository (Blake and Merz 1998) as well as the glaucoma data (see Section 5). Characteristics of the problems are given in Table 3. We draw 500 random samples from the out-of-bag performance measures (misclassification or mean-squared error) in a dependent K-sample design as described in the conceptual framework for benchmark experiments of Hothorn, Leisch, Zeileis, and Hornik (2005).

The performance of conditional inference trees is compared to the performance of exhaustive search trees with pruning (as implemented in **rpart**) and unbiased **QUEST** trees (nominal responses) and piecewise constant **GUIDE** trees (numeric responses), respectively. The tree sizes for **QUEST** and **GUIDE** are determined by pruning as well.

Two performance distributions are said to be equivalent when the performance of the conditional

	J	n	NA	m	nominal	ordinal	continuous
Boston Housing	–	506	–	13	–	–	13
Ozone	–	361	158	12	3	–	9
Servo	–	167	–	4	4	–	–
Breast Cancer	2	699	16	9	4	5	–
Diabetes	2	768	–	8	–	–	8
Glass	6	214	–	9	–	–	9
Glaucoma	2	196	–	62	–	–	62
Ionosphere	2	351	–	33	1	–	32
Sonar	2	208	–	60	–	–	60
Soybean	19	683	121	35	35	5	–
Vehicle	4	846	–	19	–	–	19
Vowel	11	990	–	10	1	–	9

Table 3: Summary of the benchmarking problems showing the number of classes of a nominal response J ('–' indicates a continuous response), the number of observations n, the number of observations with at least one missing value (NA) as well as the measurement scale and number m of the covariates.

inference trees compared to the performance of one competitor (**rpart**, **QUEST** or **GUIDE**) does not differ by an amount of more than 10%. The null hypothesis of non-equivalent performances is then defined in terms of the ratio of the expectations of the performance distribution of conditional inference trees and its competitors. Equivalence can be established at level α based on two one-sided level α tests by the intersection-union principle (Berger and Hsu 1996). Here, this corresponds to a rejection of the null hypothesis of non-equivalence performances at the 5% level when the 90% two-sided Fieller (1940) confidence interval for the ratio of the performance expectations is completely included in the equivalence range $(0.9, 1.1)$.

The boxplots of the pairwise ratios of the performance measure evaluated for conditional inference trees and pruned exhaustive search trees (**rpart**, Figure 6) and pruned unbiased trees (**QUEST**/**GUIDE**, Figure 7) are accomplished by estimates of the ratio of the expected performances and corresponding Fieller confidence intervals. For example, an estimate of the ratio of the misclassification errors of **rpart** and conditional inference trees for the glaucoma data of 1.043 means that the misclassification error of conditional inference trees is 4.3% larger than the misclassification error of **rpart**. The confidence interval of $(1.023, 1.064)$ leads to the conclusion that this inferiority is within the pre-defined equivalence margin of $\pm 10\%$ and thus the performance of conditional inference trees is on par with the performance of **rpart** for the glaucoma data.

Equivalent performance between conditional inference trees and **rpart** cannot be postulated for the Glass data. The performance of the conditional inference trees is roughly 10% worse compared with **rpart**. In all other cases, the performance of conditional inference trees is better than or equivalent to the performance of exhaustive search (**rpart**) and unbiased procedures (**QUEST** or **GUIDE**) with pruning. The conditional inference trees perform better compared to **rpart** trees by a magnitude of 25% (Boston Housing), 10% (Ionosphere) and 15% (Ozone). The improvement upon unbiased **QUEST** and piecewise constant **GUIDE** models is 10% for the Boston Housing data and 50% for the Ionosphere and Soybean data. For all other problems, the performance of conditional inference trees fitted within a permutation testing framework can be assumed to be equivalent to the performance of all three competitors.

The simulation experiments with model (4) presented in the first paragraph on estimation accuracy lead to the impression that the partitions induced by **rpart** trees are structurally different from the partition induced by conditional inference trees. Because the 'true' partition is unknown for the datasets used here, we compare the partitions obtained from conditional inference trees and **rpart** by their normalized mutual information. The median normalized mutual information is 0.447 and a bivariate density estimate depicted in Figure 8 does not indicate any relationship between the

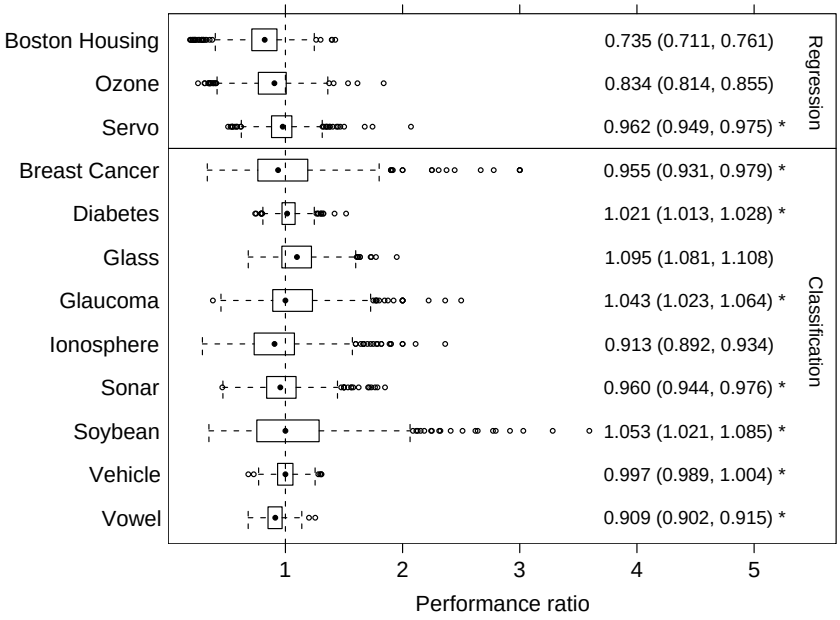

Figure 6: Distribution of the pairwise ratios of the performances of the conditional inference trees and **rpart** accomplished by estimates and 90% Fieller confidence intervals for the ratio of the expectations of the performance distributions. Stars indicate equivalent performances, i.e., the confidence interval is covered by the equivalence range $(0.9, 1.1)$.

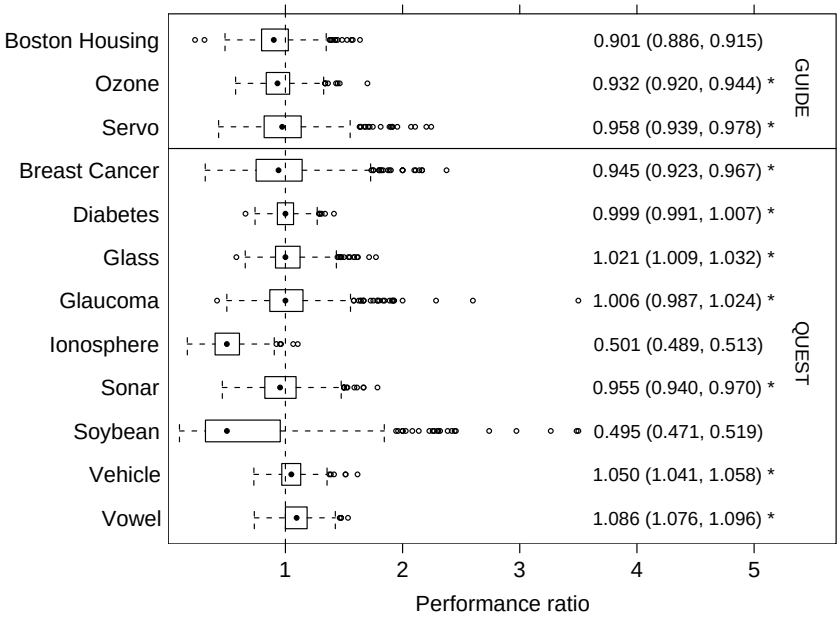

Figure 7: Distribution of the pairwise ratios of the performances of the conditional inference trees and **QUEST** (classification) or **GUIDE** (regression) accomplished by estimates and 90% Fieller confidence intervals for the ratio of the expectations of the performance distributions. Stars indicate equivalent performances, i.e., the confidence interval is covered by the equivalence range $(0.9, 1.1)$.

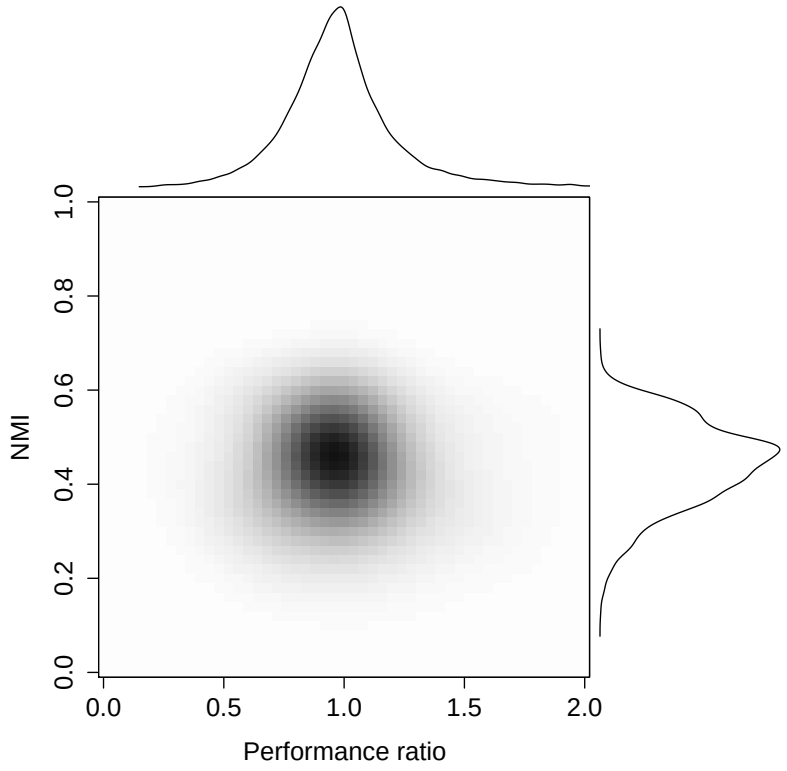

Figure 8: Distribution of the pairwise performance ratios of conditional inference trees and **rpart** and the normalized mutual information measuring the discrepancy of the induced partitions.

ratio of the performances and the discrepancy of the partitions.

This results is interesting from a practical point of view. It implies that two recursive partitioning algorithms can achieve the same prediction accuracy but, at the same time, represent structurally different regression relationships, i.e., different models and thus may lead to different conclusions about the influence of certain covariates on the response.

7. Discussion

In this paper, recursive binary partitioning with piecewise constant fits, a popular tool for regression analysis, is embedded into a well-defined framework of conditional inference procedures. Both the overfitting and variable selection problems induced by a recursive fitting procedure are solved by the application of the appropriate statistical test procedures to both variable selection and stopping. Therefore, the conditional inference trees suggested in this paper are not just heuristics but non-parametric models with well-defined theoretical background. The methodology is generally applicable to regression problems with arbitrary measurement scales of responses and covariates. In addition to its advantageous statistical properties, our framework is computationally attractive since we do not need to evaluate all $2^{K-1} - 1$ possible splits of a nominal covariate at K levels for the variable selection. In contrast to algorithms incorporating pruning based on resampling, the models suggested here can be fitted deterministically, provided that the exact conditional distribution is not approximated by Monte-Carlo methods.

The simulation and benchmarking experiments in Section 6 support two conclusions: Conditional inference trees as suggested in this paper select variables in an unbiased way and the partitions

induced by this recursive partitioning algorithm are not affected by overfitting. Even in a very simple simulation model, the partitions obtained from conditional inference trees are, on average, closer to the true data partition compared to partitions obtained from an exhaustive search procedure with pruning. When the response is independent of all covariates, the proportion of incorrect decisions in the root node is limited by α and when the response is associated with one of the covariates, conditional inference trees select the correct covariate more often than the exhaustive search procedure. In the light of these findings, the conditional inference trees seem to be more appropriate for diagnostic purposes than exhaustive search procedures. The results of the benchmarking experiments with real data show that the prediction accuracy of conditional inference trees is competitive with the prediction accuracy of both an exhaustive search procedure (**rpart**) and unbiased recursive partitioning (**QUEST**/**GUIDE**) which select the tree size by pruning. Therefore, our findings contradict the common opinion that pruning procedures outperform algorithms with internal stopping with respect to prediction accuracy. From our point of view, internal stopping criteria based on hypothesis tests evaluated earlier (see for example the results of Frank and Witten 1998) suffer from that fact that the data are transformed in order to fit the requirements of a certain test procedure, such as categorizing continuous variables for a χ^2 test, instead of choosing a test procedure defined for the original measurement scale of the covariates.

When the parameter α is interpreted as a pre-defined nominal level of the permutation tests performed in every node of the tree, the tree structures visualized in a way similar to Figures 1–3 are valid in a sense that covariates without association to the response appear in a node only with a probability not exceeding α. Moreover, subject matter scientists are most likely more familiar with the interpretation of α as pre-defined nominal level of hypothesis tests rather than as a fine-tuned hyper parameter. Although it is possible to choose α in a data-dependent way when prediction accuracy is the main focus, the empirical experiments in Section 6 show that the classical convention of $\alpha = 0.05$ performs well compared to tree models optimizing the prediction accuracy directly. However, while the predictions obtained from conditional inference trees are as good as the predictions of pruned exhaustive search trees, the partitions induced by both algorithms differ structurally. Therefore, the interpretations obtained from conditional inference trees and trees fitted by an exhaustive search without bias correction cannot be assumed to be equivalent. Thus, two rather different partitions, and therefore models, may have equal prediction accuracy. Since a key reason for the popularity of tree based methods stems from their ability to represent the estimated regression relationship in an intuitive way, interpretations drawn from regression trees must be taken with a grain of salt.

In summary, this paper introduces a statistical approach to recursive partitioning. Formal hypothesis tests for both variable selection and stopping criterion are established. This choice leads to tree structured regression models for all kinds of regression problems, including models for censored, ordinal or multivariate response variables. Because well-known concepts are the basis of variable selection and stopping criterion, the resulting models are easier to communicate to practitioners. Simulation and benchmark experiments indicate that conditional inference trees are well-suited for both explanation and prediction.

Acknowledgements

We would like to thank three anonymous referees, one associate editor and the editor of JCGS for their valuable comments which lead to substantial improvements. The work of T. Hothorn was supported by Deutsche Forschungsgemeinschaft (DFG) under grant HO 3242/1-1.

References

Agresti A (2002). *Categorical Data Analysis*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition.

Berger RL, Hsu JC (1996). "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets." *Statistical Science*, **11**(4), 283–319. With discussion.

Blake C, Merz C (1998). "UCI Repository of Machine Learning Databases." URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

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

De'ath G (2002). "Multivariate Regression Trees: A New Technique For Modeling Species-Environment Relationships." *Ecology*, **83**(4), 1105–1117.

Dobra A, Gehrke J (2001). "Bias Correction in Classification Tree Construction." In "Proceedings of the Eighteenth International Conference on Machine Learning," pp. 90–97. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1.

Fieller EC (1940). "The Biological Standardization of Insulin." *Journal of the Royal Statistical Society, Supplement*, **7**, 1–64.

Frank E, Witten IH (1998). "Using a Permutation Test for Attribute Selection in Decision Trees." In "Proceedings of the Fifteenth International Conference on Machine Learning," pp. 152–160. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

Genz A (1992). "Numerical Computation of Multivariate Normal Probabilities." *Journal of Computational and Graphical Statistics*, **1**, 141–149.

Goodman LA (1965). "On Simultaneous Confidence Intervals for Multinomial Proportions." *Technometrics*, **7**(2), 247–254.

Hosmer DW, Lemeshow S (2000). *Applied Logistic Regression*. John Wiley & Sons, New York, 2nd edition.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006). "A Lego System for Conditional Inference." *The American Statistician*, **60**, 257–263. doi:10.1198/000313006X118430.

Hothorn T, Leisch F, Zeileis A, Hornik K (2005). "The Design and Analysis of Benchmark Experiments." *Journal of Computational and Graphical Statistics*, **14**(3), 675–699.

Jensen DD, Cohen PR (2000). "Multiple Comparisons in Induction Algorithms." *Machine Learning*, **38**, 309–338.

Kass G (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Applied Statistics*, **29**(2), 119–127.

Kim H, Loh WY (2001). "Classification Trees With Unbiased Multiway Splits." *Journal of the American Statistical Association*, **96**(454), 589–604.

Kim H, Loh WY (2003). "Classification Trees with Bivariate Linear Discriminant Node Models." *Journal of Computational and Graphical Statistics*, **12**, 512–530.

Lausen B, Hothorn T, Bretz F, Schumacher M (2004). "Assessment of Optimal Selected Prognostic Factors." *Biometrical Journal*, **46**(3), 364–374.

Lausen B, Schumacher M (1992). "Maximally Selected Rank Statistics." *Biometrics*, **48**, 73–85.

LeBlanc M, Crowley J (1992). "Relative Risk Trees for Censored Survival Data." *Biometrics*, **48**, 411–425.

LeBlanc M, Crowley J (1993). "Survival Trees by Goodness of Split." *Journal of the American Statistical Association*, **88**(422), 457–467.

Loh WY (2002). "Regression Trees With Unbiased Variable Selection And Interaction Detection." *Statistica Sinica*, **12**, 361–386.

Loh WY, Shih YS (1997). "Split Selection Methods for Classification Trees." *Statistica Sinica*, **7**, 815–840.

Loh WY, Vanichsetakul N (1988). "Tree-Structured Classification via Generalized Discriminant Analysis." *Journal of the American Statistical Association*, **83**, 715–725. With discussion.

Mardin CY, Hothorn T, Peters A, Jünemann AG, Nguyen NX, Lausen B (2003). "New Glaucoma Classification Method Based on Standard HRT Parameters by Bagging Classification Trees." *Journal of Glaucoma*, **12**(4), 340–346.

Martin JK (1997). "An Exact Probability Metric for Decision Tree Splitting and Stopping." *Machine Learning*, **28**, 257–291.

Miller R, Siegmund D (1982). "Maximally Selected Chi Square Statistics." *Biometrics*, **38**, 1011–1016.

Mingers J (1987). "Expert Systems – Rule Induction with Statistical Data." *Journal of the Operations Research Society*, **38**(1), 39–47.

Molinaro AM, Dudoit S, van der Laan MJ (2004). "Tree-based Multivariate Regression and Density Estimation with Right-Censored Data." *Journal of Multivariate Analysis*, **90**(1), 154–177.

Morgan JN, Sonquist JA (1963). "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association*, **58**, 415–434.

Murthy SK (1998). "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey." *Data Mining and Knowledge Discovery*, **2**, 345–389.

Noh HG, Song MS, Park SH (2004). "An Unbiased Method for Constructing Multilabel Classification Trees." *Computational Statistics & Data Analysis*, **47**(1), 149–164.

O'Brien SM (2004). "Cutpoint Selection for Categorizing a Continuous Predictor." *Biometrics*, **60**, 504–509.

Peters A, Hothorn T, Lausen B (2002). "ipred: Improved Predictors." *R News*, **2**(2), 33–36. ISSN 1609-3631, URL http://CRAN.R-project.org/doc/Rnews/.

Quinlan JR (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, California.

Rasch D (1995). *Mathematische Statistik*. Johann Ambrosius Barth Verlag, Heidelberg, Leipzig.

R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org/.

Schumacher M, Holländer N, Schwarzer G, Sauerbrei W (2001). "Prognostic Factor Studies." In J Crowley (ed.), "Statistics in Clinical Oncology," pp. 321–378. Marcel Dekker, New York, Basel.

Segal MR (1988). "Regression Trees for Censored Data." *Biometrics*, **44**, 35–47.

Shih YS (1999). "Families of Splitting Criteria for Classification Trees." *Statistics and Computing*, **9**, 309–315.

Shih YS (2004). "A Note on Split Selection Bias in Classification Trees." *Computational Statistics & Data Analysis*, **45**, 457–466.

Strasser H, Weber C (1999). "On the Asymptotic Theory of Permutation Statistics." *Mathematical Methods of Statistics*, **8**, 220–250. URL http://epub.wu-wien.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_94c.

Strehl A, Ghosh J (2003). "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions." *Journal of Machine Learning Research*, **3**, 583–617.

Therneau TM, Atkinson EJ (1997). "An Introduction to Recursive Partitioning using the rpart Routine." *Technical Report 61*, Section of Biostatistics, Mayo Clinic, Rochester. URL http://www.mayo.edu/hsr/techrpt/61.pdf.

Westfall PH, Young SS (1993). *Resampling-based Multiple Testing*. John Wiley & Sons, New York.

White AP, Liu WZ (1994). "Bias in Information-based Measures in Decision Tree Induction." *Machine Learning*, **15**, 321–329.

Zhang H (1998). "Classification Trees for Multiple Binary Responses." *Journal of the American Statistical Association*, **93**, 180–193.

Appendix A

An equivalent but computational simpler formulation of the linear statistic for case weights greater than one can be written as follows. Let $\mathbf{a} = (a_1, \ldots, a_{\mathbf{w}_\cdot})$, $a_l \in \{1, \ldots, n\}, l = 1, \ldots, \mathbf{w}_\cdot$, denote the vector of observation indices, with index i occuring w_i times. Instead of recycling the ith observation w_i times it is sufficient to implement the index vector $\mathbf{a}$ into the computation of the test statistic and its expectation and covariance. For one permutation σ of $\{1, \ldots, \mathbf{w}_\cdot\}$, the linear statistic (1) may be written as

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = \text{vec}\left(\sum_{k=1}^{\mathbf{w}_\cdot} g_j(X_{ja_k}) h(\mathbf{Y}_{\sigma(\mathbf{a})_k}, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n))^\top\right) \in \mathbb{R}^{p_j q}$$

now taking case weights greater zero into account.

Appendix B

The results shown in Section 5 are, up to some labelling, reproducible using the following R code:

```
library("party")

data("GlaucomaM", package = "ipred")
plot(ctree(Class ~ ., data = GlaucomaM))

data("GBSG2", package = "ipred")
plot(ctree(Surv(time, cens) ~ ., data = GBSG2))

data("mammoexp", package = "party")
plot(ctree(ME ~ ., data = mammoexp))
```