Unbiased Recursive Partitioning I:  
A Non-parametric Conditional Inference Framework

Torsten Hothorn\textsuperscript{1}, Kurt Hornik\textsuperscript{2} and Achim Zeileis\textsuperscript{2}

\textsuperscript{1} Friedrich-Alexander-Universität Erlangen-Nürnberg  
\textsuperscript{2} Wirtschaftsuniversität Wien
Recursive Partitioning


Many variants have been (and still are) published, the majority of which are special cases of a simple two-stage algorithm:

**Step 1:** partition the observations by univariate splits in a recursive way

**Step 2:** fit a constant model in each cell of the resulting partition.

Two Fundamental Problems

**Overfitting:** 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.

**Selection Bias:** Exhaustive search methods suffer from a severe selection bias towards covariates with many possible splits and / or missing values.
A Statistical Approach

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.

and present a unified framework embedding recursive binary partitioning into the well
defined theory of

**Part I:** permutation tests developed by Strasser and Weber (1999),

**Part II:** tests for parameter instability in (parametric) regression models.
The Regression Problem

The distribution of a (possibly multivariate) response $Y \in \mathcal{Y}$ is to be regressed on a $m$-dimensional covariate vector $X = (X_1, \ldots, X_m) \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$:

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

based on a learning sample of $n$ observations

$$\mathcal{L}_n = \{(Y_i, X_{1i}, \ldots, X_{mi}); i = 1, \ldots, n\}.$$

possibly with case counts $w = (w_1, \ldots, w_n)$. We are interested in estimating $f$. 

A Generic Algorithm

1. For case weights $w$ 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 covariate $X_{j^*}$ with strongest association to $Y$.

2. Choose a set $A^* \subset X_{j^*}$ in order to split $X_{j^*}$ into two disjoint sets $A^*$ and $X_{j^*} \setminus A^*$. The case weights $w_{\text{left}}$ and $w_{\text{right}}$ 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 $w_{\text{left}}$ and $w_{\text{right}}$, respectively.
Recursive Partitioning by Conditional Inference

In each node identified by case weights $w$, the global hypothesis of independence is formulated in terms of the $m$ partial hypotheses $H_0^j : D(Y|X_j) = D(Y)$ with global null hypothesis $H_0 = \bigcap_{j=1}^{m} H_0^j$.

Stop recursion when $H_0$ can not be rejected at a pre-specified level $\alpha$.

Else: Measure the association between $Y$ and each of the covariates $X_j, j = 1, \ldots, m$, by test statistics or $P$-values indicating the deviation from the partial hypotheses $H_0^j$. 
Linear Statistics

Use a (multivariate) linear statistic

$$T_j(L_n, w) = \text{vec} \left( \sum_{i=1}^{n} w_i g_j(X_{ji}) h(Y_i, (Y_1, \ldots, Y_n))^\top \right) \in \mathbb{R}^{p_j q}$$

with $g_j$ being a transformation of $X_j$ and influence function $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$. This type of statistics was suggested by Strasser and Weber (1999).
Conditional Expectation and Covariance under $H^j_0$

\[
\mu_j = \mathbb{E}(T_j(\mathcal{L}_n, w) | S(\mathcal{L}_n, w)) = \text{vec} \left( \left( \sum_{i=1}^{n} w_i g_j(X_{ji}) \right) \mathbb{E}(h | S(\mathcal{L}_n, w))^\top \right),
\]

\[
\Sigma_j = \mathbb{V}(T_j(\mathcal{L}_n, w) | S(\mathcal{L}_n, w)) = \frac{w}{w \cdot 1} \mathbb{V}(h | S(\mathcal{L}_n, w)) \otimes \left( \sum_i w_i g_j(X_{ji}) \otimes w_i g_j(X_{ji})^\top \right)
\]

\[
- \frac{1}{w \cdot 1} \mathbb{V}(h | S(\mathcal{L}_n, w)) \otimes \left( \sum_i w_i g_j(X_{ji}) \right) \otimes \left( \sum_i w_i g_j(X_{ji}) \right)^\top,
\]

with $w \cdot = \sum_{i=1}^{n} w_i$. 
Test Statistics

A (multivariate) linear statistic $T_j$ can now be used to construct a test statistic for testing $H_0^j$, for example via

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

or

$$c_{\text{quad}}(t, \mu, \Sigma) = (t - \mu)\Sigma^+(t - \mu)^\top$$
Variable Selection and Stopping Criteria

Test $H_0$ based on $P_1, \ldots, P_m,$

$$P_j = \mathbb{P}_{H_0}^{j}(c(T_j(L_n, w), \mu_j, \Sigma_j) \geq c(t_j, \mu_j, \Sigma_j)|S(L_n, w))$$

conditional on all permutations $S(L_n, w)$ of the data. This solves the overfitting problem.

When we can reject $H_0$ in step 1 of the generic algorithm we select the covariate with minimum $P$-value

$$P_j = \mathbb{P}_{H_0}^{j}(c(T_j(L_n, w), \mu_j, \Sigma_j) \geq c(t_j, \mu_j, \Sigma_j)|S(L_n, w))$$

of the conditional test for $H_0^j$. This prevents a variable selection bias.
Splitting Criteria

The goodness of a split is evaluated by two-sample linear statistics which are special cases of the linear statistic $\mathbf{T}$. For all possible subsets $A$ of the sample space $\mathcal{X}_j^*$ the linear statistic

$$T^A_j(\mathcal{L}_n, \mathbf{w}) = \text{vec} \left( \sum_{i=1}^{n} w_i I(X_{j^*i} \in A) h(Y_i, (Y_1, \ldots, Y_n))^\top \right)$$

induces a two-sample statistic and we implement the best split

$$A^* = \arg\max_A c(t^A_{j^*}, \mu^A_{j^*}, \Sigma^A_{j^*}).$$
Examples: Glaucoma & 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. For 98 patients and 98 controls, matched by age and gender, 62 covariates describing the eye morphology are available.

```
R> data(GlaucomaM, package = "ipred")
R> gtree <- ctree(Class ~ ., data = GlaucomaM)
R> plot(gtree)
```
Node 4 (n = 51)

Node 5 (n = 22)

Node 6 (n = 14)

Node 7 (n = 109)
Examples: Glaucoma & Laser Scanning Images

Interested in the class distribution in each *inner* node? Want to explore the process of the split statistics in each inner node?

```R
R> plot(gtree, inner_panel = node_barplot, tnex = 1)
```
Examples: Glaucoma & Laser Scanning Images

R> cex <- 1.6
R> inner <- nodes(gtree, 1:3)
R> layout(matrix(1:length(inner), ncol = length(inner)))
R> out <- sapply(inner, function(i) {
+   splitstat <- i$psplit$splitstatistic
+   x <- GlaucomaM[[i$psplit$variableName]][splitstat > 0]
+   plot(x, splitstat[splitstat > 0], main = paste("Node", i$nodeID), xlab = i$psplit$variableName, ylab = "Statistic", ylim = c(0, 10), cex.axis = cex, cex.lab = cex, cex.main = cex)
+   abline(v = i$psplit$splitpoint, lty = 3)
+ })
Examples: Node Positive Breast Cancer

Evaluation of prognostic factors for the German Breast Cancer Study Group (GBSG2) data, a prospective controlled clinical trial on the treatment of node positive breast cancer patients. Complete data of seven prognostic factors of 686 women are used for prognostic modeling

\[ R> \text{data(GBSG2, package = } "\text{ipred}"\text{)} \]
\[ R> \text{stree } \leftarrow \text{ctree(Surv(time, cens) } \sim ., \text{ data = GBSG2)} \]
\[ R> \text{plot(stree)} \]
Examples: 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.

\begin{verbatim}
R> data(mammoexp, package = "party")
R> mtree <- ctree(ME ~ ., data = mammoexp, scores = list(ME = 1:3, + SYMPT = 1:4, DECT = 1:3))
R> plot(mtree)
\end{verbatim}
SYMPT $p < 0.001$

$\leq$ Agree $\leq$ Agree

Node 2 (n = 113)

Never Within a Year Over a Year

0
0.2
0.4
0.6
0.8
1

Node 4 (n = 208)

$\leq 8$ $> 8$

Node 5 (n = 91)
Benchmark Experiments

**Hypothesis 1:** Conditional inference trees with statistical stop criterion perform as good as an exhaustive search algorithm with pruning.

**Hypothesis 2:** Conditional inference trees with statistical stop criterion perform as good as parametric unbiased recursive partitioning (QUEST, GUIDE, Loh, 2002, is a starting point).

Equivalence measured by ratio of misclassification or mean squared error with a equivalence margin of $\pm 10\%$ and Fieller confidence intervals.
ctree vs. rpart

Performance ratio

Boston Housing 0.735 (0.711, 0.761)
Breast Cancer 0.955 (0.931, 0.979) *
Diabetes 1.021 (1.013, 1.028) *
Glass 1.095 (1.081, 1.108)
Glaucoma 1.043 (1.023, 1.064) *
Ionosphere 0.913 (0.892, 0.934)
Ozone 0.834 (0.814, 0.855)
Servo 0.962 (0.949, 0.975) *
Sonar 0.960 (0.944, 0.976) *
Soybean 1.053 (1.021, 1.085) *
Vehicle 0.997 (0.989, 1.004) *
Vowel 0.909 (0.902, 0.915) *
ctree vs. QUEST / GUIDE

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Performance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston Housing</td>
<td>0.901 (0.886, 0.915)</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>0.945 (0.923, 0.967) *</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.999 (0.991, 1.007) *</td>
</tr>
<tr>
<td>Glass</td>
<td>1.021 (1.009, 1.032) *</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>1.006 (0.987, 1.024) *</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>0.501 (0.489, 0.513)</td>
</tr>
<tr>
<td>Ozone</td>
<td>0.932 (0.920, 0.944) *</td>
</tr>
<tr>
<td>Servo</td>
<td>0.958 (0.939, 0.978) *</td>
</tr>
<tr>
<td>Sonar</td>
<td>0.955 (0.940, 0.970) *</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.495 (0.471, 0.519)</td>
</tr>
<tr>
<td>Vehicle</td>
<td>1.050 (1.041, 1.058) *</td>
</tr>
<tr>
<td>Vowel</td>
<td>1.086 (1.076, 1.096) *</td>
</tr>
</tbody>
</table>

Performance ratio ranges are given with 95% confidence intervals.
Summary

The separation of variable selection and split point estimation first implemented in ‘CHAID’ (Kass, 1980) is the basis for unbiased recursive partitioning for responses and covariates measured at arbitrary scales.

The statistical internal stop criterion ensures that interpretations drawn from such trees are valid in a statistical sense, i.e., with appropriate control of type I errors.

Even the algorithm has no concept of prediction error, the performance is at least equivalent to established procedures.

We are committed to reproducible research, see

\[R > \text{vignette(package = "party")}\]


