# Generic Frameworks for Nonparametric and Parametric Model Trees

Achim Zeileis, Torsten Hothorn, Kurt Hornik

http://eeecon.uibk.ac.at/~zeileis/

# Overview

- Motivation: Trees, leaves, and branches
- Conditional inference trees
  - Conditional inference
  - Splitting and pruning
- Model-based recursive partitioning
  - Model estimation
  - Tests for parameter instability
  - Segmentation
  - Pruning
- Applications
  - Pima Indians diabetes
  - Treatment effect for chronic disease
- Software
- Summary

# Motivation: Trees

Breiman (2001, *Statistical Science*) distinguishes two cultures of statistical modeling.

- **Data models:** Stochastic models, typically parametric.
- **Algorithmic models:** Flexible models, data-generating process unknown.

**Example:** Recursive partitioning models dependent variable $Y$ by "learning" a partition w.r.t explanatory variables $Z_1, \dots, Z_l$.

**Key features**:

- Predictive power in nonlinear regression relationships.
- Interpretability (enhanced by visualization), i.e., no "black box" methods.

# Motivation: Leaves

**Typically:** Simple models for univariate $Y$, e.g., mean or proportion.

**Examples**: CART and C4.5 in statistical and machine learning, respectively.

**Problems:** For classical tree algorithms.

- No concept of "significance", possibly biased variable selection.
- No complex (parametric) models in leaves.
- Many different tree algorithms for different types of data.

**Solutions:** Flexible generic frameworks based on statistical inference.

- Nonparametric: Employ only empirical distribution for inference.
- Parametric: Synthesis of parametric data models and algorithmic tree models.

# Motivation: Branches

**Base algorithm**: Growth of branches from the roots to the leaves of the tree typically follows a simple *recursive partitioning* algorithm.

1. Fit a (possibly very simple) model for the response $Y$.
2. Assess association of $Y$ and each $Z_j$.
3. Split sample along the $Z_{j^*}$ with strongest association: Choose breakpoint with highest improvement of the model fit.
4. Repeat steps 1–3 recursively in the subsamples until some stopping criterion is met.

**Generally:** Tree algorithms differ w.r.t. choice of model (1), association measure (2), split strategy (3) and stopping criterion or “pruning” strategy (4).

# Conditional inference trees

**Idea:** Fully nonparametric approach using a modern framework unifying classical nonparametric tests.

**Algorithm**

1. Model: Nonparametric, empirical distribution of $Y$.
2. Association measure: Permutation test (i.e., conditional inference) for independence of $Y$ and each $Z_j$.
3. Split strategy: Maximize two-smaple contrast of $Y$ along $Z_j^*$.
4. Stopping criterion: Significance of test in step 2.

**Note:** Both model and tests condition on the observed data.

# Conditional inference trees

**Model:** Predictions can be computed from any quantity of the empirical distribution of $Y$ in the relevant node, e.g., the mean/median/etc. for numeric $Y$, proportion of “successes” for binary $Y$, Kaplan-Meier survivor function for censored $Y$, etc.

**Association measure:** Independence tests derived from general correlation of $Y$ and $Z_j$.

$$t_j \quad = \quad \mathrm{vec} \left( \sum_{i=1}^{n} h(Y_i) \cdot g(Z_{j,i}) \right),$$

with $p$-dimensional transformation $g(\cdot)$ and $q$-dimensional influence function $h(\cdot)$.

# Conditional inference trees

**Test statistics:** Scalar standardized statistic based on conditional expectation $\mu_j$ and covariance matrix $\Sigma_j$ (given the data).

$$
\begin{aligned}
s_{\max}(t, \mu, \Sigma) &= \max_k \left| \frac{(t-\mu)_k}{\sqrt{\Sigma_{k,k}}} \right|, \\
s_{\text{quad}}(t, \mu, \Sigma) &= (t-\mu)\Sigma^{+}(t-\mu).
\end{aligned}
$$

Under independence, all permutations of $Y$ yield the conditional distribution of $t_j$. Taking expectations w.r.t. this yields:

$$
\begin{aligned}
\mu_j = \mathsf{E}(t_j) &= \operatorname{vec}\left( \left( \sum_{i=1}^{n} g(Z_{j,i}) \right) \mathsf{E}(h)^\top \right), \\
\mathsf{E}(h) &= n^{-1} \sum_i h(Y_i),
\end{aligned}
$$

# Conditional inference trees

**Similarly:** $pq \times pq$ conditional covariance matrix $\Sigma_j$ computed from the permutation distribution under independence:

$$\begin{aligned}
\Sigma_j = \mathrm{Var}(t_j) &= \frac{n}{n-1}\mathrm{Var}(h) \otimes \left(\sum_i g(Z_{j,i}) \otimes g(Z_{j,i})^\top\right) - \\
&\quad \frac{1}{n-1}\mathrm{Var}(h) \otimes \left(\sum_i g(Z_{j,i})\right) \otimes \left(\sum_i g(Z_{j,i})\right)^\top, \\
\mathrm{Var}(h) &= n^{-1} \sum_i \left(h(Y_i) - \mathrm{E}(h)\right)\left(h(Y_i) - \mathrm{E}(h)\right)^\top,
\end{aligned}$$

where $\otimes$ denotes the Kronecker product.

# Conditional inference trees

**Significance:** Various approaches can be used to assess the significance of the test statistic $s(t_j, \mu_j, \Sigma_j)$:

- Exact: Direct computation of the statistic for all permutations. Typically burdensome.
- Approximate: Compute statistics for a sufficiently large number of permutations, drawn using Monte Carlo methods.
- Asymptotic: Compute the conditional asymptotic distribution of $s$ based on the asymptotic conditional distribution of $t_j$.
  $t_j \sim \mathcal{N}(\mu_j, \Sigma_j)$.

# Conditional inference trees

**Choice of transformations:** Based on scale of Y and $Z_j$ and type of dependence.

- Categorical: Indicator functions for all $C$ categories $h(y) = (I_1(y), \dots, I_C(y))^\top$.
- Numeric:
  - Location: $h(y) = y$ or $h(y) = \text{rank}(y)$.
  - Scatter: $h(y) = (y - \bar{y})^2$ or $h(y) = (\text{rank}(y) - (n+1)/2)^2$.
  - Threshold: $h(y) = I(y > \zeta)$.
- Survival: Log rank scores

**Special cases:** Choice of $h(\cdot)$ and analogously $g(\cdot)$ yields many classical tests as special cases. Wilcoxon-Mann-Whitney, Spearman, Pearson's $\chi^2$, Cochran-Armitage, log rank, Kruskal-Wallis, and many more.

# Conditional inference trees

**Split strategy:** Maximize two-smaple contrast of $Y$ along $Z_j$.

- Employ threshold transformation $g(Z_j) = I(Z_j > \zeta)$ for all possible thresholds $\zeta$.
- Choose split $\zeta^*$ that maximizes the associated test statistic.

**Stopping criterion:** Non-significance of Bonferroni-adjusted $p$ values from permutation tests.

# Application: Pima Indians diabetes

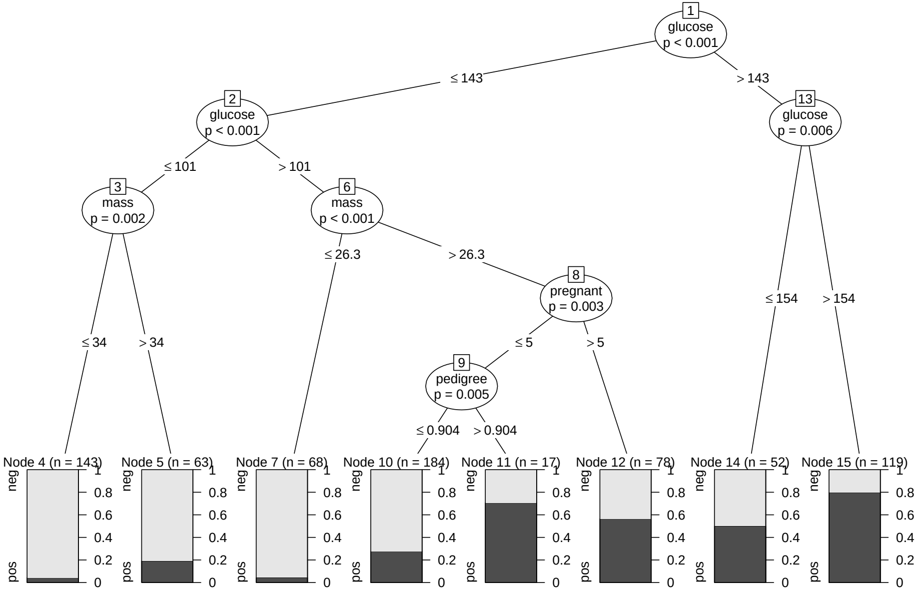
**Task:** Classification of diabetes in Pima Indian women.

**Source:** Asuncion & Newman (2007), UCI Repository of Machine Learning Databases.
`http://www.ics.uci.edu/~mlearn/MLRepository.html`.

**Response:** Test result for *diabetes* (positive/negative).

**Explanatory variables:** Plasma *glucose* concentration, number of times *pregnant*, diastolic blood *pressure* (mm Hg), body *mass* index, diabetes *pedigree* function, *age* (in years).

# Application: Pima Indians diabetes



# Application: Pima Indians diabetes

**Inference:** In each node, an asymptotic permutation test for independence of *diabetes* ($Y$) and each of the variables *glucose*, ..., *age* ($Z_1, \dots, Z_6$) is carried out.

**Transformations:** Indicator function for categorical variables (response) and identity for numeric variables (all regressors).

$$\begin{aligned} h(Y_i) &= (I_{\text{pos}}(Y_i), I_{\text{neg}}(Y_i))^\top, \\ g(Z_{j,i}) &= Z_{j,i}. \end{aligned}$$

**Interpretation:** Corresponds to two-sample $t$ test with pooled one-sample standard deviation.

# Model-based recursive partitioning

**Idea:** More complex models for multivariate $Y$, e.g., multivariate normal model, regression models, etc.

**Goal:**

- Synthesis of parametric data models and algorithmic tree models.
- Fitting local models by partitioning of the sample space.

**Algorithm**

1. Model: Parametric model for $Y$ with additive objective function.
2. Association measure: Parameter instability tests.
3. Split strategy: Model segmentation.
4. Stopping criterion: Significance of test in step 2.

# Model-based recursive partitioning: Estimation

**Models:** $\mathcal{M}(Y, \theta)$ with (potentially) multivariate observations $Y \in \mathcal{Y}$ and $k$-dimensional parameter vector $\theta \in \Theta$.

**Parameter estimation:** $\widehat{\theta}$ by optimization of objective function $\Psi(Y, \theta)$ for $n$ observations $Y_i$ ($i = 1, \dots, n$):

$$\widehat{\theta} \quad = \quad \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi(Y_i, \theta).$$

**Special cases:** Maximum likelihood (ML), weighted and ordinary least squares (OLS and WLS), quasi-ML, and other M-estimators.

**Central limit theorem:** If there is a true parameter $\theta_0$ and given certain weak regularity conditions, $\hat{\theta}$ is asymptotically normal with mean $\theta_0$ and sandwich-type covariance.

# Model-based recursive partitioning: Estimation

**Estimating function:** $\widehat{\theta}$ can also be defined in terms of

$$\sum_{i=1}^{n} \psi(Y_i, \widehat{\theta}) = 0,$$

where $\psi(Y, \theta) = \partial \Psi(Y, \theta) / \partial \theta$.

**Idea:** In many situations, a single global model $\mathcal{M}(Y, \theta)$ that fits **all** $n$ observations cannot be found. But it might be possible to find a partition w.r.t. the variables $Z = (Z_1, \dots, Z_l)$ so that a well-fitting model can be found locally in each cell of the partition.

**Tool:** Assess parameter instability w.r.t to partitioning variables $Z_j \in \mathcal{Z}_j$ ($j = 1, \dots, l$).

# Model-based recursive partitioning: Tests

Generalized M-fluctuation tests capture instabilities in $\widehat{\theta}$ for an ordering w.r.t $Z_j$.

**Basis:** Empirical fluctuation process of cumulative deviations w.r.t. to an ordering $\sigma(Z_{ij})$.

$$W_j(t, \widehat{\theta}) \quad = \quad \widehat{V}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \psi(Y_{\sigma(Z_{ij})}, \widehat{\theta}) \qquad (0 \leq t \leq 1)$$

**Functional central limit theorem:** Under parameter stability $W_j(\cdot, \widehat{\theta}) \stackrel{d}{\longrightarrow} W^0(\cdot)$, where $W^0$ is a $k$-dimensional Brownian bridge.

# Model-based recursive partitioning: Tests

**Test statistics:** Scalar functional $\lambda(W_j)$ that captures deviations from zero.

**Null distribution:** Asymptotic distribution of $\lambda(W^0)$.

**Special cases:** Class of test encompasses many well-known tests for different classes of models. Certain functionals $\lambda$ are particularly intuitive for numeric and categorical $Z_j$, respectively.

**Advantage:** Model $\mathcal{M}(Y, \widehat{\theta})$ just has to be estimated once. Empirical estimating functions $\psi(Y_i, \widehat{\theta})$ just have to be re-ordered and aggregated for each $Z_j$.

# Model-based recursive partitioning: Tests

**Splitting numeric variables:** Assess instability using sup*LM* statistics.

$$\lambda_{\sup LM}(W_j) \quad = \quad \max_{i=\underline{i},\dots,\bar{\imath}} \left(\frac{i}{n} \cdot \frac{n-i}{n}\right)^{-1} \left\| W_j\left(\frac{i}{n}\right)\right\|_2^2 .$$

**Interpretation:** Maximization of single shift *LM* statistics for all conceivable breakpoints in $[\underline{i}, \bar{\imath}]$.

**Limiting distribution:** Supremum of a squared, *k*-dimensional tied-down Bessel process.

# Model-based recursive partitioning: Tests

**Splitting categorical variables:** Assess instability using $\chi^2$ statistics.

$$\lambda_{\chi^2}(W_j) \quad = \quad \sum_{c=1}^{C} \frac{n}{|I_c|} \left\| \Delta_{I_c} W_j \left( \frac{i}{n} \right) \right\|_2^2$$

**Feature:** Invariant for re-ordering of the $C$ categories and the observations within each category.

**Interpretation:** Captures instability for split-up into $C$ categories.

**Limiting distribution:** $\chi^2$ with $k \cdot (C - 1)$ degrees of freedom.

# Model-based recursive partitioning: Segmentation

**Goal:** Split model into $b = 1, \dots, B$ segments along the partitioning variable $Z_j$ associated with the highest parameter instability. Local optimization of

$$\sum_{b} \sum_{i \in I_b} \Psi(Y_i, \theta_b).$$

$B = 2$: Exhaustive search of order $O(n)$.

$B > 2$: Exhaustive search is of order $O(n^{B-1})$, but can be replaced by dynamic programming of order $O(n^2)$. Different methods (e.g., information criteria) can choose $B$ adaptively.

**Here:** Binary partitioning.

# Model-based recursive partitioning: Pruning

**Pruning:** Avoid overfitting.

**Pre-pruning:** Internal stopping criterion. Stop splitting when there is no significant parameter instability.

**Post-pruning:** Grow large tree and prune splits that do not improve the model fit (e.g., via cross-validation or information criteria).

**Here:** Pre-pruning based on Bonferroni-corrected $p$ values of the fluctuation tests.

# Application: Pima Indians diabetes

**Task:** Reconsider classification of diabetes in Pima Indian women.

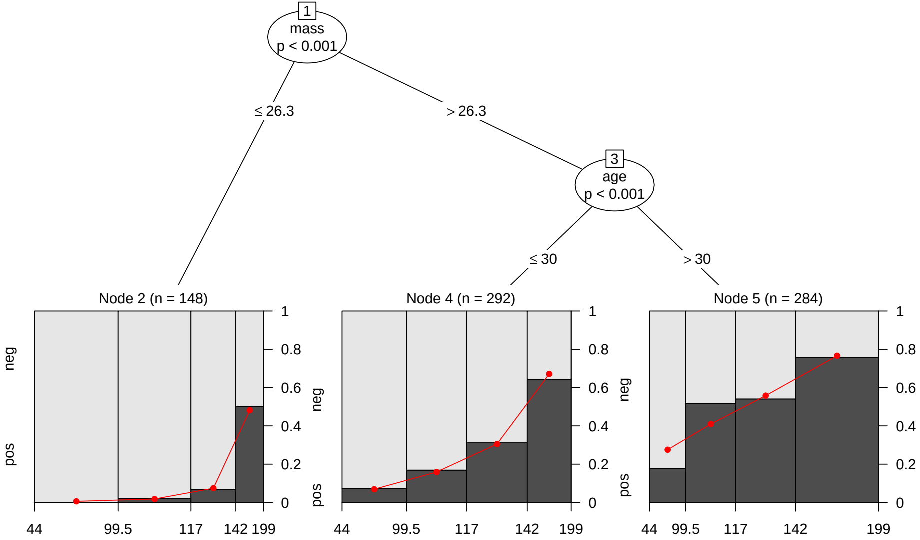
**Idea:**

- Variable *glucose* occurred in many splits in conditional inference tree.
- More parsimonious model may be possible if *glucose* is employed as continuous regressor rather than partitioning variable.

**Model:** Logistic regression of *diabetes* on *glucose*.

**Partitioning variables:** All remaining variables.

# Application: Pima Indians diabetes



# Application: Pima Indians diabetes

**Model-based recursive partitioning:**

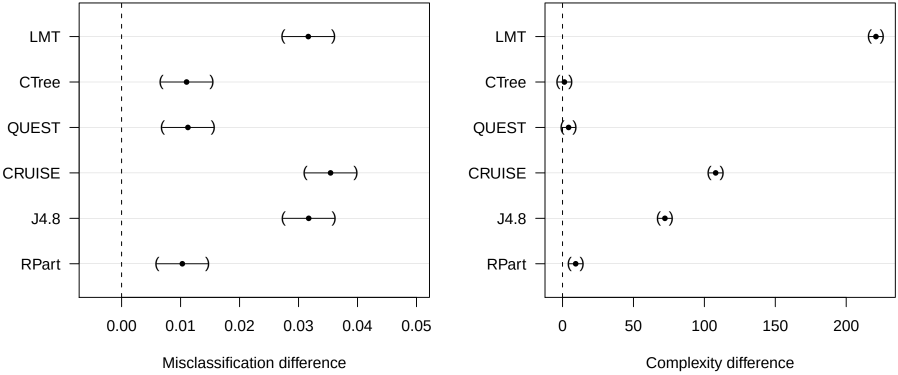
- Coefficient estimates for regressors.
- Parameter instability tests for partitioning variables (bold = significant at adjusted 5% level, underlined = smallest $p$ value).

| | Regressors | | Partitioning variables | | | | |
|---|---|---|---|---|---|---|---|
| | (const.) | glucose | pregnant | pressure | mass | pedigree | age |
| 1 | −5.608 | 0.039 | **26.49** | 8.67 | **<u>43.41</u>** | **21.04** | **39.47** |
| 2 | −10.999 | 0.065 | 8.40 | 4.50 | <u>9.31</u> | 4.02 | 4.53 |
| 3 | −4.958 | 0.037 | **24.80** | 7.63 | 9.05 | **19.29** | **<u>33.71</u>** |
| 4 | −6.573 | 0.045 | 3.46 | 3.77 | 5.09 | <u>7.20</u> | 6.20 |
| 5 | −3.319 | 0.027 | 6.24 | 1.74 | 13.34 | <u>14.89</u> | 10.24 |

# Application: Pima Indians diabetes

**Benchmark:** Compare predictive performance (misclassification rate) and model complexity (number of parameters/splits) of model-based recursive partitioning with other tree algorithms.

**Setup:** 250 bootstrap samples and out-of-bag misclassification rate.



# Application: Treatment effect for chronic disease

**Task:** Identify groups of chronic disease patients with different treatment effects.

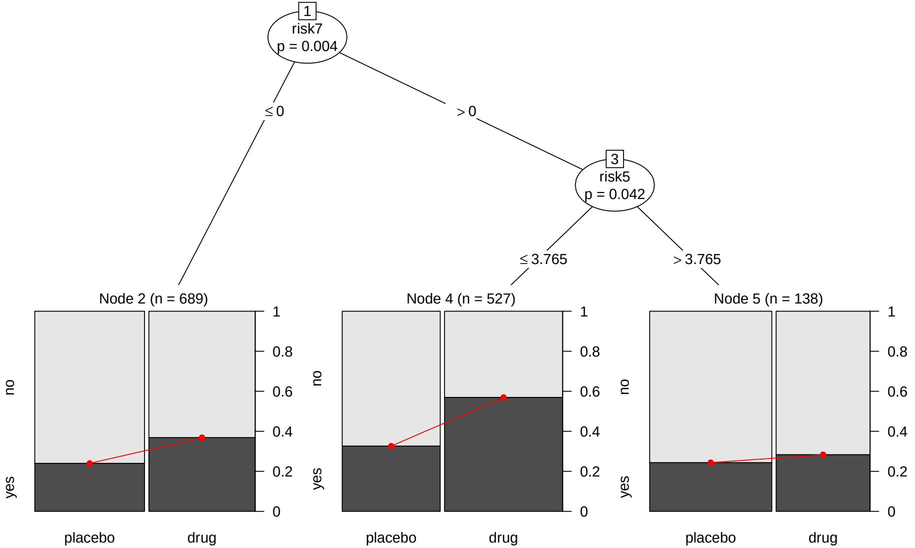
**Source:** Anonymized data from consulting project.

**Model:** Logistic regression.

- Response: Improvement (yes/no) of chronic disease after treatment over several weeks.
- Regressor: Treatment (active drug/placebo).
- Partitioning variables: 11 variables that describe disease status of patients. Lower values indicate more severe forms of the disease.

**Result:** Treatment is most effective for certain intermediate forms of the disease.

# Application: Treatment effect for chronic disease



# Software

All methods are implemented in the R system for statistical computing and graphics. Freely available under the GPL (General Public License) from the Comprehensive R Archive Network:

- Trees/recursive partytioning: `ctree()` in **party** for conditional inference trees, and `mob()` in **party** for model-based recursive partitioning.
- Inference: `independence_test()` in **coin** for permutation tests for independence, and `gefp()` in **strucchange** for structural change tests.

`http://www.R-project.org/`
`http://CRAN.R-project.org/`

# Summary

**Conditional inference trees:**

- Tree models based on nonparametric statistical inference.
- Based on modern class of permutation tests for independence.
- Aims to capture dependence patterns by recursive partitioning.
- Can be adapted to dependent and explanatory variables of arbitrary types, by employing suitable transformations/influence functions.
- Flexible implementation freely available: New transformations/influence functions can be simply plugged in.

# Summary

**Model-based recursive partitioning:**

- Synthesis of classical parametric data models and algorithmic tree models.
- Based on modern class of parameter instability tests.
- Aims to minimize clearly defined objective function by greedy forward search.
- Can be applied general class of parametric models.
- Alternative to traditional means of model specification, especially for variables with unknown association.
- Object-oriented implementation freely available: Extension for new models requires some coding but not too extensive if interfaced model is well designed.

# References


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

Zeileis A, Hornik K (2007). "Generalized M-Fluctuation Tests for Parameter Instability." *Statistica Neerlandica*, **61**(4), 488–508. doi:10.1111/j.1467-9574.2007.00371.x

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674. doi:10.1198/106186006X133933

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