Model-based Recursive Partitioning

Achim Zeileis Torsten Hothorn Kurt Hornik

http://statmath.wu-wien.ac.at/~zeileis/

Overview

- Motivation
- The recursive partitioning algorithm
  - Model fitting
  - Testing for parameter instability
  - Splitting
  - Pruning
- Illustrations
  - Demand for economic journals
  - Pima Indians diabetes
- Summary

Motivation

Starting point: Most recursive partitioning algorithms learn a partition/segmentation from data and then fit a naive model in each terminal node, e.g., a mean, relative frequencies or a Kaplan-Meier curve.

Idea: Employ parametric models in each node. Solutions exist only for special cases, e.g., linear regression (M5', GUIDE), logistic regression (LMT).

Goal: Unified algorithm for constructing general segmented parametric models by recursive partitioning.

Parametric models

Consider models $\mathcal{M}(Y, \theta)$ with (possibly vector-valued) observations $Y \in \mathcal{Y}$ and a k-dimensional vector of parameters $\theta \in \Theta$.

Given n observations Y_i ($i = 1, \dots, n$) the model can be fit by minimizing some objective function $\Psi(Y, \theta)$ yielding the parameter estimate $\widehat{\theta}$

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

This type of estimators includes maximum likelihood (ML), ordinary least squares (OLS), Quasi-ML and further M-type estimators.

Segmented models

Idea: In many situations, it is unreasonable to assume that a single global model $\mathcal{M}(Y, \theta)$ can be fit to **all** n observations. But it might be possible to partition the observations with respect to covariates $Z = (Z_1, \dots, Z_l)$ such that a fitting model can be found in each cell of the partition.

Goal: Learn partition via recursive partitioning with respect to $Z_j \in \mathcal{Z}_j$ $(j = 1, \dots, l)$.

The recursive partitioning algorithm

1. Fit the model once to all observations in the current node by estimating $\widehat{\theta}$ via minimization of Ψ.
2. Assess whether the parameter estimates are stable with respect to every ordering $Z_1, \dots, Z_l$. If there is some overall instability, select the variable Z_j associated with the highest parameter instability, otherwise stop.
3. Compute the split point(s) that locally optimize Ψ (either for a fixed number of splits, or choose the number of splits adaptively).
4. Split this node into daughter nodes and repeat the procedure.

1. Model fitting

Under mild regularity conditions it can be shown that the estimate $\widehat{\theta}$ can also be computed by solving the first order conditions

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

where

$$\psi(Y, \theta) = \frac{\partial \Psi(Y, \theta)}{\partial \theta}$$

is the score function or estimating function corresponding to $\Psi(Y, \theta)$.

2. Testing for parameter instability

Generalized M-fluctuation tests (Zeileis & Hornik, 2003) can be used to assess whether the parameter estimates $\widehat{\theta}$ are stable over a certain variable or not.

Capture instabilities in an empirical fluctuation process of cumulative scores for each ordering of the observations

$$W(t, \widehat{\theta}) \quad = \quad \widehat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \psi(Y_i, \widehat{\theta}) \qquad (0 \leq t \leq 1)$$

and assess its fluctuation by a suitable functional.

Assessing numerical variables

The most intuitive functional for assessing the stability with respect to a numerical partitioning variable Z_j is the supLM statistic of Andrews (1993):

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

This gives the maximum of the single changepoint LM statistics over all possible changepoints in $[\underline{i}, \overline{\imath}]$.

The limiting distribution is given by the supremum of a squared, k-dimensional tied-down Bessel process.

Assessing categorical variables

To assess the stability of a categorical variable with C levels, a χ^2 statistic is most intuitive

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

because it is insensitive to re-ordering of the levels and the observations within the levels.

It essentially captures the instability when splitting the model into C groups.

The limiting distribution is χ^2 with $k \cdot (C - 1)$ degrees of freedom.

3. Splitting

A single optimal split of the observations with respect to Z_j into $B = 2$ partitions can easily be computed in $O(n)$ by exhaustive search.

For $B > 2$, when an exhaustive search would be of order $O(n^{B-1})$, the optimal partition can be found using a dynamic programming approach of order $O(n^2)$ (Hawkins, 2001; Bai & Perron, 2003) or via iterative algorithms (Muggeo, 2003).

Various algorithms for adaptively choosing the number of segments B are available, e.g., via information criteria.

Pruning

The algorithm described so far employs a **pre-pruning** strategy, i.e., uses an internal stopping criterion: if no variable exhibits significant parameter instability, the algorithm stops.

Alternatively/additionally, a **post-pruning** strategy can be used. This seems particularly attractive if ML is used for parameter estimation. Then a ML tree can be grown which is consequently associated with a segmented ML model. This can be pruned afterwards using information criteria for example.

Example: Demand for econ. journals

Goal: Explain demand for economic journals (number of library subscriptions in logs).

Clear: Demand depends on price (price per citation, also in logs)

Here: Segment the demand equation, a linear regression, with respect to further variables such as age, number of characters, society etc.

Example: Demand for econ. journals

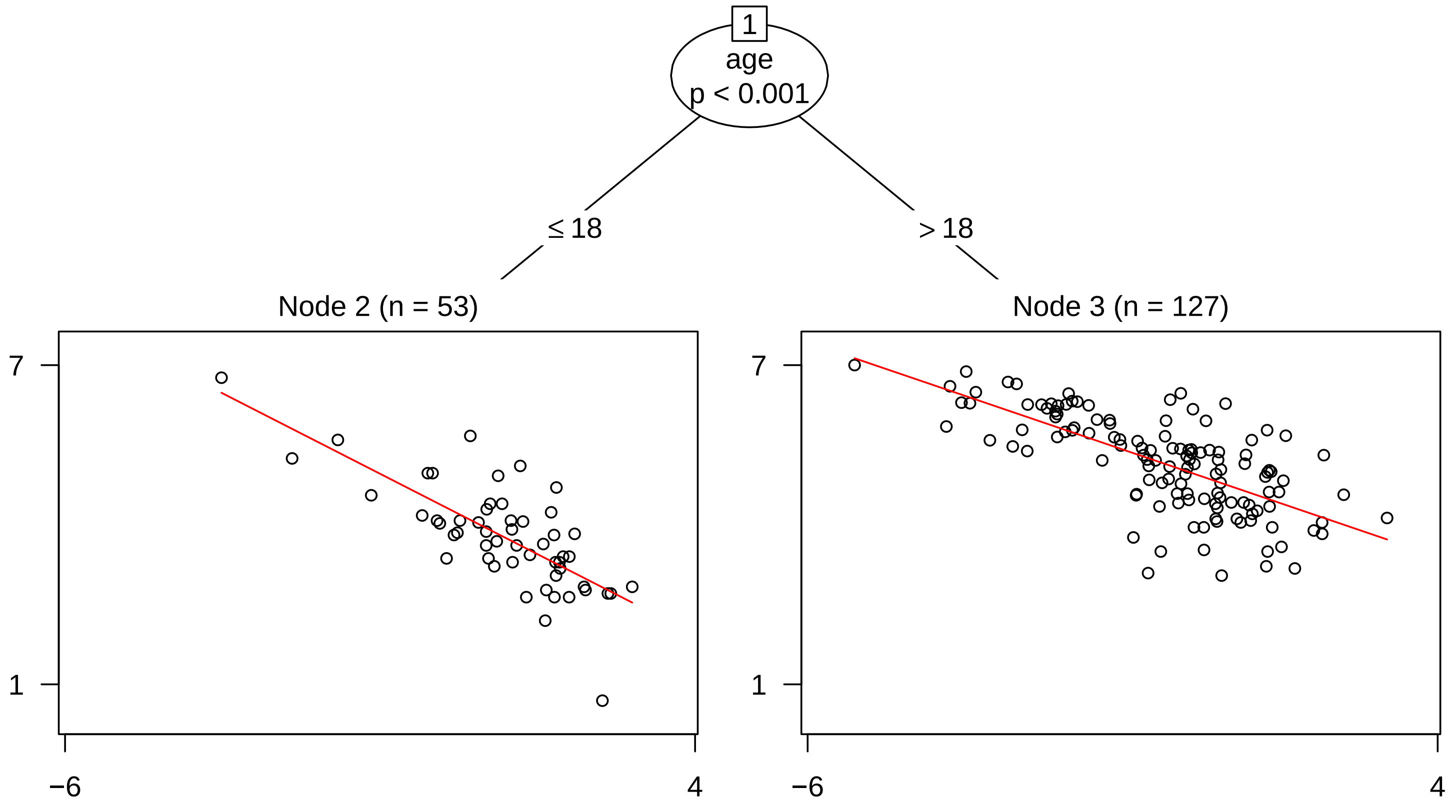
```
R> fmJ <- mob(subs ~ citeprice | society + citations + age + chars + price,
+    data = journals, model = linearModel, control = mob_control(minsplit = 10))

-------------------------------------------
Fluctuation tests of splitting variables:
            society citations          age    chars      price
statistic 3.2797248 5.2614434 4.219816e+01 4.563841 16.3127521
p.value   0.6598605 0.9958892 1.465145e-07 0.999475  0.0489191

Best splitting variable: age
Perform split? yes
-------------------------------------------
Node properties:
age <= 18; criterion = 1, statistic = 42.198
...

R> plot(fmJ)
```

Example: Demand for econ. journals



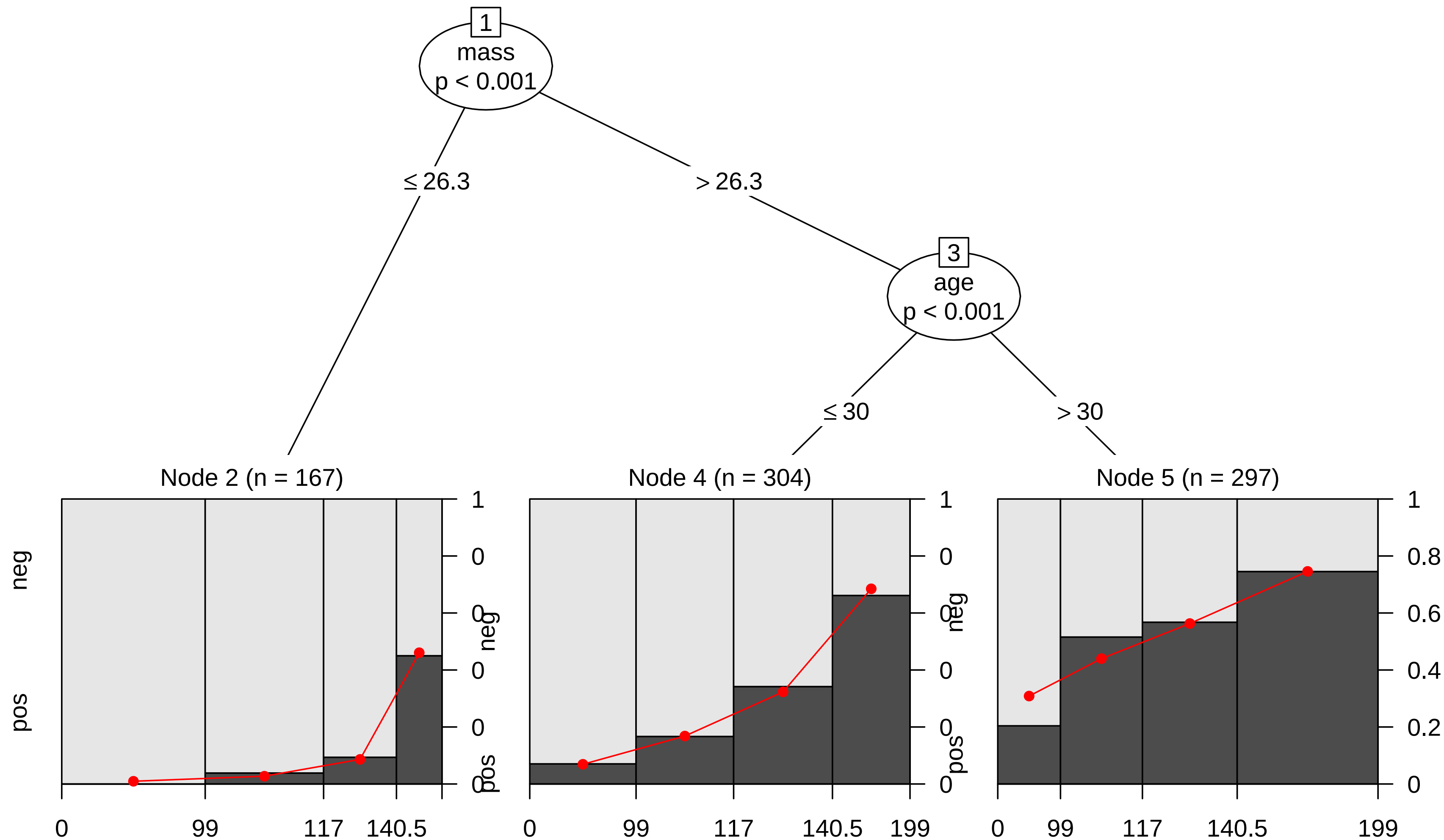
Example: Pima Indians diabetes

Goal: Explain outcome of a test for diabetes among Pima Indian women.

Clear: Outcome depends on plasma glucose concentration.

Here: Segment a logistic regression with explanatory variable glucose. All remaining variables are used as partitioning variables.

Example: Pima Indians diabetes



Summary

Model-based recursive partitioning:

- based on well-established statistical models,
- aims at minimizing a clearly defined objective function (and not certain heuristics),
- unbiased due to separation of variable and cutpoint selection,
- statistically motivated stopping criterion,
- employs general class of tests for parameter instability.
- available in function `mob()` in package **party** available from

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

References

Andrews DWK (1993). "Tests for Parameter Instability and Structural Change With Unknown Change Point." *Econometrica*, **61**, 821–856.

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), Forthcoming.

Zeileis A, Hornik K (2003). "Generalized M-Fluctuation Tests for Parameter Instability." *Report 80*, SFB "Adaptive Information Systems and Modelling in Economics and Management Science". URL http://www.wu-wien.ac.at/am/reports.htm#80.

Zeileis A, Hothorn T, Hornik K (2005). "Model-based Recursive Partitioning." *Report 19*, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL http://epub.wu-wien.ac.at/.

Zeileis A, Hothorn T, Hornik K (2006). "Evaluating Model-based Trees in Practice." *Report 32*, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL http://epub.wu-wien.ac.at/.