Model-Based Boosting: Unbiased Variable Selection and Model Choice

Benjamin Hofner [1]

Institut für Medizininformatik, Biometrie und Epidemiologie (IMBE)

Friedrich-Alexander-Universität Erlangen-Nürnberg

joint work with

Torsten Hothorn, Thomas Kneib and Matthias Schmid

DAGStat 2010 - Statistik unter einem Dach

24.03.2010

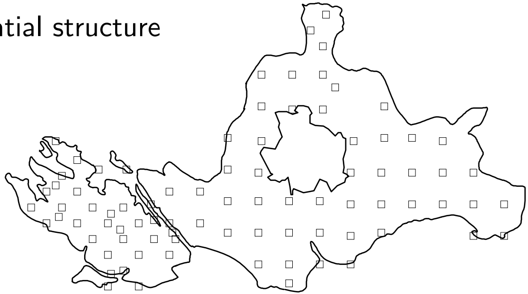
TU Dortmund

[1] benjamin.hofner@imbe.med.uni-erlangen.de

Forest Health Data

- **Aim:** Identify predictors of the health status of trees
- **Data:** Yearly visual forest health inventories carried out from 1983 to 2004 in a northern Bavarian forest district (Spessart)
- 83 plots of beeches within a 15 km $\times$ 10 km area
- **Response:** binary defoliation indicator at plot i in year t: ($y_{it} = 1$ defoliation above 25%)
- Large data set ($n = 1793$)

$\Rightarrow$ Longitudinal data with spatial structure



Covariates

Continuous:
- average age of trees at the observation plot
- elevation above sea level in meters
- inclination of slope in percent
- depth of soil layer in centimeters
- pH-value at 0-2cm depth
- density of forest canopy in percent

Categorical:
- thickness of humus layer in 5 ordered categories
- base saturation in 4 ordered categories

Binary:
- type of stand
- application of fertilisation

- Previous analyses resulted in models that contained linear and smooth effects as well as categorical covariates.
- Additionally, a spatial effect and a random effect for the plot could be identified.

$\Rightarrow$ **Boosting can estimate all effects and includes intrinsic variable selection and model choice.**

Model Fitting with Component-Wise Boosting

Structured Additive Model

$$\mu_i = \mathbb{E}(y|\mathbf{x}_i) = h(\eta_i(\mathbf{x}_i))$$

with response function h and additive predictor

$$\eta_i(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{J} f_j(\mathbf{x}_i),$$

- Model fitting aims at **minimizing the expected loss** with appropriate loss function ρ, e.g.,
  - squared error loss $\rho(y, \eta(\mathbf{x})) = (y - \eta(\mathbf{x}))^2$ for Gaussian models
  - negative log-likelihood for GLMs
- In practice: Minimization of the **empirical risk**

$$n^{-1} \sum_{i=1}^{n} \rho(y_i, \eta_i(\mathbf{x}_i))$$

Boosting

- minimizes empirical risk (e.g., negative log likelihood)
- in a stagewise fashion
- via functional gradient descent (FGD).

In each iteration m

- (negative) gradient of the loss function $u_i^{[m]} = -\left.\frac{\partial \rho(y_i, \eta)}{\partial \eta}\right|_{\eta = \hat{\eta}_i^{[m-1]}}$ is estimated via base-learners ($\hat{\mathbf{u}}^{[m]} = \hat{g}_j(\mathbf{x})$)
- update only model term corresponding to the best-fitting base-learner $\hat{g}_{j^*}$ (based on the RSS):
  add a **small fraction** ν **of the estimate** $\hat{g}_{j^*}$ (e.g., 10%) to the model

$\Rightarrow$ variable and model selection is achieved

Practical notes

- Base-learners represent functions $f_j(\cdot)$ from structured additive predictor (in the simplest case)
- We get an interpretable model similar to models from MLE
- Regularization via base-learner selection and shrinkage

Problems (and a Solution)

- **Variable selection** and **model choice** can be **seriously biased** if some base-learners offer higher flexibility.
  - Variable Selection Bias:
    e.g., categorical covariate (with many categories) $\succ$ continuous covariate
  - Model Choice Bias:
    e.g., smooth effect $\succ$ linear effect
- Unbiased (or at least improved) selection desired

- **Possible solution:** Make the competitors comparable with respect to their flexibility (measured by the degrees of freedom)

Problems (and a Solution)

- **Variable selection** and **model choice** can be **seriously biased** if some base-learners offer higher flexibility.
  - Variable Selection Bias:
    e.g., categorical covariate (with many categories) $\succ$ continuous covariate
  - Model Choice Bias:
    e.g., smooth effect $\succ$ linear effect
- Unbiased (or at least improved) selection desired

- **Possible solution:** Make the competitors comparable with respect to their flexibility (measured by the degrees of freedom)

Penalized Least Squares Base-Learners

Consider (penalized) least squares base-learners

$$\hat{g}_j(\mathbf{x}) = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{K})^{-1}\mathbf{X}^\top}_{=:\mathbf{S}\ \text{(smoother matrix)}}\mathbf{u}^{[m]},$$

where $\mathbf{X}$ is a suitable design matrix.

Examples of penalized LS base-learners

- Unpenalized base-learners ($\lambda = 0$)
- Ridge-penalized base-learners for unordered categorical covariates ($\mathbf{X}$ e.g., dummy coded)
- Base-learners with first order difference penalty for ordered categorical covariates (Gertheiss & Tutz, 2009)
  ($\mathbf{X}$ e.g., dummy coded)
- P-spline base-learners with second order difference penalty for continuous covariates
  ($\mathbf{X}$ B-spline basis expansion)

Penalized Least Squares Base-Learners

Central Idea

Set $\mathrm{df} = 1$ for all base-learners to prevent selection bias

NB: Final model can adopt (much) higher flexibility due to the iterative nature of boosting!

Theoretical Considerations (Hofner, Hothorn, Kneib, & Schmid, 2009)

Instead of

$$\mathrm{df} := \mathrm{trace}(\mathbf{S})$$

define

$$\mathrm{df} := \mathrm{trace}(2\mathbf{S} - \mathbf{S}^\top \mathbf{S})$$

(tailored for the comparison of RSS (see also Buja, Hastie, & Tibshirani, 1989))

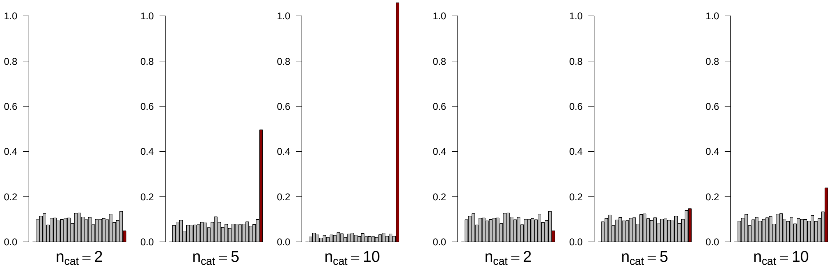
"Null Model" with Non-Informative Factor

- 25 non-informative continuous covariates
- 1 non-informative categorical covariate with increasing # of categories
- $y \sim N(0, 1)$
- $n = 150$, $B = 1000$

"Null Model" with Non-Informative Factor

- 25 non-informative continuous covariates
- 1 non-informative categorical covariate with increasing # of categories
- $y \sim N(0, 1)$
- $n = 150$, $B = 1000$

Selection Frequencies



(c) Unpenalized Base-Learner

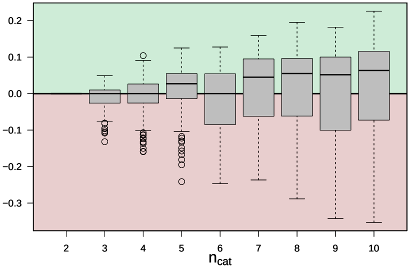
(d) Ridge Penalized Base-Learner

“Power Case” with Non-Informative Factor

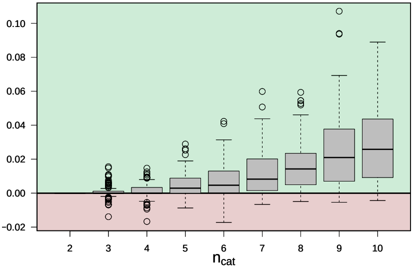
- 5 continuous covariates with $\boldsymbol{\beta}_{\text{info}} = (-2, -1, 1, 2, 3)^\top$
  20 additional non-informative continuous covariates
- 1 non-informative categorical covariate with increasing # of categories
- $y|x \sim N(x^\top \beta, \sigma^2)$, with σ^2 such that $R^2 \approx 0.3$
- $n = 150$, $B = 100$

"Power Case" with Non-Informative Factor

- 5 continuous covariates with $\boldsymbol{\beta}_{\text{info}} = (-2, -1, 1, 2, 3)^\top$
  20 additional non-informative continuous covariates
- 1 non-informative categorical covariate with increasing # of categories
- $y|x \sim N(x^\top \beta, \sigma^2)$, with σ^2 such that $R^2 \approx 0.3$
- $n = 150,\ B = 100$



(g) Difference of Relative Selection Frequencies (unpenalized - penalized)



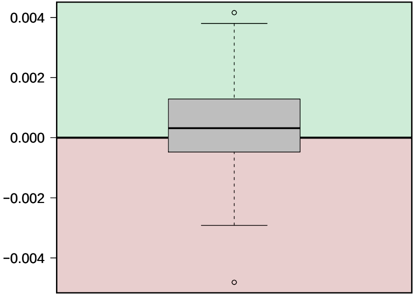
(h) $\text{MSE}_{\text{unpenalized}} - \text{MSE}_{\text{penalized}}$

"Power Case" with (Potentially) Smooth Effects

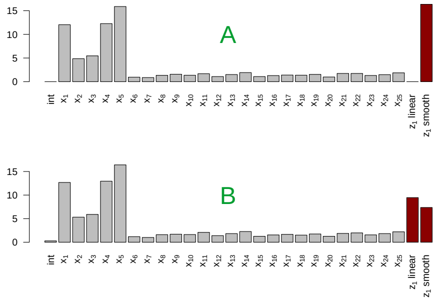
- 5 continuous covariates with $\beta_{\text{info}} = (-2, -1, 1, 2, 3)^\top$
  20 additional non-informative continuous covariates
- 1 continuous covariate with linear effect ($\beta_{z_1} = 1.5$)
- Otherwise same simulation setting as in "factor case"
- Add (A) linear effect + smooth effect (4 df)
  or (B) linear effect + smooth deviation from linearity (1 df)

"Power Case" with (Potentially) Smooth Effects

- 5 continuous covariates with $\beta_{\text{info}} = (-2, -1, 1, 2, 3)^\top$
  20 additional non-informative continuous covariates
- 1 continuous covariate with linear effect ($\beta_{z_1} = 1.5$)
- Otherwise same simulation setting as in "factor case"
- Add (A) linear effect + smooth effect (4 df)
  or (B) linear effect + smooth deviation from linearity (1 df)



(k) Partial Deviation (A - B)



(l) Selection Frequency

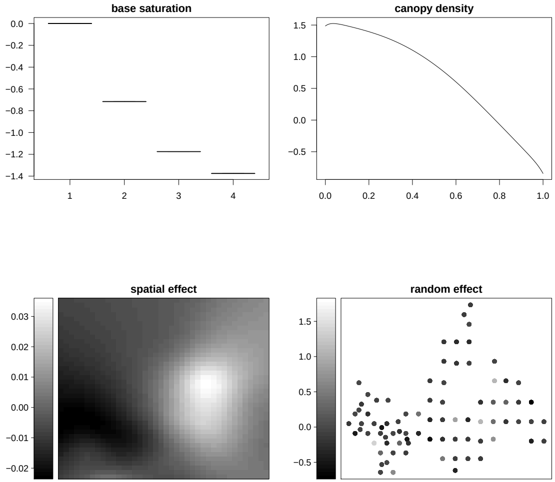
Forest Health Data - Results

Using component-wise (penalized) least squares base-learners with 1 df each, we get a **final model** with

- Parametric effects: for fertilisation (binary), base saturation (ordinal), age and calender time
- Nonparametric effect: for canopy density
- Spatial effect: + unstructured random effect (with a clear domination of the latter)

Not selected: thickness of humus layer, ph-value, soil depth, type of stand, inclination of slope, elevation above sea level

Forest Health Data - Results (ctd.)



Further linear effects

Covariates	β
Fertilization	-0.760
Age	0.016
Year	0.068

Take-Home Messages

- One can fit a wide range of models by boosting:
  (generalized) linear models, survival models, ...
  (generalized) additive models, structured additive models, ...
- Boosting results in interpretable models if one uses linear or smooth base-learners (i.e., no tree base-learners).
- Boosting (intrinsically) allows for variable / model selection.
- We get a **severe reduction** of selection bias by using penalized base-learners with equal df.
- Use a suitable definition of degrees of freedom $\mathrm{df} = \mathrm{trace}(2\mathbf{S} - \mathbf{S}^\top \mathbf{S})$.

R-package **mboost** available on CRAN to fit all the models covered in this talk (and many more) (Hothorn, Bühlmann, Kneib, Schmid, & Hofner, 2010)

References

Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *The Annals of Statistics*, *17*, 453–555.

Gertheiss, J., & Tutz, G. (2009). Penalized regression with ordinal predictors. *International Statistical Review*, *77*, 345–365.

Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2009). *A framework for unbiased model selection based on boosting* (Tech. Rep. No. 72). Department of Statistics, Ludwig-Maximilans-Universität München.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2010). *mboost: Model-based boosting.* (R package version 2.0-3)

Find out more: `http://benjaminhofner.de/`