

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

Institutskolloquium - Institut für Statistik
LMU München - 2009

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

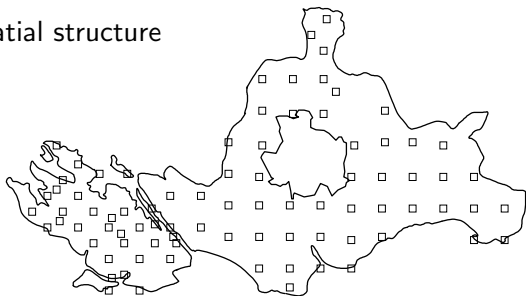
Aims

- Fitting models for (potentially) **high-dimensional data sets**
 - Good prediction performance should be achieved
 - Resulting models should be **interpretable**
 - Only relevant covariates should be included (i.e., **variable selection**)
 - Appropriate modeling alternative should be used (linear vs. flexible vs. ...) (i.e., **model choice**)
- ⇒ One solution to all of this: **Component-Wise Boosting**

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 × 10 km area
- Large data set ($n = 1793$)
- Response: binary defoliation indicator at plot i in year t : ($y_{it} = 1$ defoliation above 25%)

⇒ 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 **categorical covariates**, **linear** and **smooth effects**.
- Additionally, a **spatial effect** and a **random effect** for the plot could be identified.

⇒ **Boosting can estimate all effects and has an intrinsic variable selection and model choice.**

Problems (and Solutions)

- **Variable selection** and **model choice** can be **seriously biased**
 - **Variable Selection Bias:**
e.g., continuous covariate vs. categorical covariate (with many categories)
 - **Model Choice Bias:**
e.g., linear effect vs. smooth effect
- Unbiased (or at least improved) selection desired
- We will present **possible solutions**

Introduction

Aim in many statistical settings:

Minimize expected loss $f^* := \operatorname{argmin}_{f(\cdot)} \mathbb{E}_{Y, \mathbf{x}} (\rho(y, f(\mathbf{x})))$

Examples:

- linear regression: $\rho(y, f(\mathbf{x})) = (y - \mathbf{x}^\top \beta)^2$ [squared error loss]
- GLMs: $\rho(y, f(\mathbf{x})) = -\ell(y, f(\mathbf{x}))$ [negative log-likelihood]

Data situation:

- Response $\mathbf{y} = (y_1, \dots, y_n)^\top$
- Predictor vector $\mathbf{x}_i = (x_{1i}, \dots, x_{pi})$, $i = 1, \dots, n$
- Minimize empirical risk

$$\hat{f}^* = \operatorname{argmin}_{f(\cdot)} \sum_{i=1}^n \rho(y_i, f(\mathbf{x}_i))$$

Introduction

Aim in many statistical settings:

Minimize expected loss $f^* := \operatorname{argmin}_{f(\cdot)} \mathbb{E}_{Y, \mathbf{x}} (\rho(y, f(\mathbf{x})))$

Examples:

- linear regression: $\rho(y, f(\mathbf{x})) = (y - \mathbf{x}^\top \beta)^2$ [squared error loss]
- GLMs: $\rho(y, f(\mathbf{x})) = -\ell(y, f(\mathbf{x}))$ [negative log-likelihood]

Data situation:

- Response $\mathbf{y} = (y_1, \dots, y_n)^\top$
- Predictor vector $\mathbf{x}_i = (x_{1i}, \dots, x_{pi})$, $i = 1, \dots, n$
- Minimize empirical risk

$$\hat{f}^* = \operatorname{argmin}_{f(\cdot)} \sum_{i=1}^n \rho(y_i, f(\mathbf{x}_i))$$

(Functional) Gradient Descent

We want to estimate f^*

⇒ Minimize empirical risk

⇒ Use iterative algorithm, e.g., gradient descent method:

- Initialize estimator $\hat{f}^{[0]}$ with offset value
- Update:

$$\hat{f}^{[m]} = \hat{f}^{[m-1]} + \nu \cdot \underbrace{\left(-\frac{\partial \rho}{\partial f}(y, \hat{f}^{[m-1]}) \right)}_{\text{negative gradient}}$$

(with step-length factor ν)

To include covariates:

Restrict f to parametric functions of \mathbf{x}_i , i.e., minimize empirical risk subject to $f \in$ parametric class

Functional Gradient Descent Boosting

- ① **Initialization:** $m := 0$; Initialize estimate with offset value

$$\hat{f}^{[0]}(\cdot) \equiv \operatorname{argmin}_c \frac{1}{n} \sum_{i=1}^n \rho(y_i, c)$$

- ② **Negative gradient:** $m := m + 1$;

$$u_i^{[m]} = - \left. \frac{\partial \rho(y_i, f)}{\partial f} \right|_{f=\hat{f}^{[m-1]}(\mathbf{x}_i)}, \quad i = 1, \dots, n$$

- ③ **Estimation:** Fit the negative gradient vector $\mathbf{u}^{[m]} = (u_1^{[m]}, \dots, u_n^{[m]})$ to $\mathbf{x}_1, \dots, \mathbf{x}_n$ by a real-valued base-learner $\hat{g}^{[m]}(\cdot)$ (for example a penalized linear model)

- ④ **Update:** Compute the update for the estimate

$$\hat{f}^{[m]}(\cdot) = \hat{f}^{[m-1]}(\cdot) + \nu \cdot \hat{g}^{[m]}(\cdot)$$

with step-length factor $0 < \nu \leq 1$.

- ⑤ **Stopping rule:** Continue iterating steps (2) to (4) until $m = m_{\text{stop}}$ for a given stopping iteration m_{stop} .

Remarks

- Base-learner $\hat{g}^{[m]}(\cdot)$ can be regarded as an approximation of the negative gradient vector $\mathbf{u}^{[m]}$
- As $\hat{g}^{[m]}(\cdot)$ is a “statistical model” (e.g., a linear model) we get estimates $\hat{f}^{[m_{\text{stop}}]}$ from the same “statistical model”:

$$\hat{f}^{[m_{\text{stop}}]}(\cdot) = \hat{f}^{[0]}(\cdot) + \sum_{m=1}^{m_{\text{stop}}} \nu \cdot \hat{g}^{[m]}(\cdot)$$

- if $\hat{g}^{[m]}(\cdot)$ is a linear model, we get a linear model (LM) for $\hat{f}^{[m_{\text{stop}}]}$
 - if $\hat{g}^{[m]}(\cdot)$ is an additive model, we get an additive model (GAM) for $\hat{f}^{[m_{\text{stop}}]}$
 - ...
- Step-length ν is “no real tuning parameter” but governs how fast the algorithm converges (as long as it is small enough)
- **Major tuning parameter:** m_{stop}
(choose $\hat{m}_{\text{stop,opt}}$ via cross validation, bootstrap, ...)
- If we use the **L2 loss** (“linear regression case”), the negative gradient reduces to least squares residuals as in a LM.
 \Rightarrow Boosting can be regarded as refitting residuals.

Structured Additive Models

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

with response function h and **additive** predictor

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

Generic representation of covariate effects $f_j(\mathbf{x}_i)$

a) **linear effects**: $f_j(\mathbf{x}_i) = f_{j,\text{linear}}(\tilde{x}_i) = \tilde{x}_i\beta$

b) **smooth effects**: $f_j(\mathbf{x}_i) = f_{j,\text{smooth}}(\tilde{x}_i)$

c) **categorical effects**: $f_j(\mathbf{x}_i) = \tilde{\mathbf{z}}_i^\top \boldsymbol{\gamma}$

d) further effects as

spatial effects, random effects, varying coefficients, ...

where \tilde{x}_i is an element of the vector \mathbf{x}_i (and $\tilde{\mathbf{z}}_i$ a corresponding dummy-coded categorical covariate).

P-Splines

flexible terms can be represented using P-splines (Eilers & Marx, 1996)

- model term:

$$f_{j,\text{smooth}}(\tilde{x}_i) = \sum_{m=1}^M \beta_{jm} B_{jm}(\tilde{x}_i)$$

- penalty:

$$\text{pen}_j(\beta_j) = \lambda_j \beta_j' \mathbf{K} \beta_j$$

with

- $\mathbf{K} = \mathbf{D}'\mathbf{D}$ (i.e., cross product of difference matrix \mathbf{D})

$$\mathbf{D} \stackrel{\text{e.g.}}{=} \begin{pmatrix} 1 & -2 & 1 & \dots & \\ 0 & 1 & -2 & 1 & \dots \end{pmatrix} \quad (\text{here differences of order 2})$$

- λ_j smoothing parameter
(larger $\lambda_j \Rightarrow$ more penalization \Rightarrow smoother fit)

Component-Wise Boosting

For variable selection and model choice, we use **component-wise** boosting.

- specify a **separate base-learner** for each covariate
(= **variable selection**)
- possible extension: specify a separate base-learner **for each modeling alternative** (e.g., linear effect vs. smooth effect)
(= **model choice**)
- base-learners represent functions $f_j(\cdot)$ from structured additive predictor
- **update** only the **best-fitting base-learner** in each step

Component-Wise Functional Gradient Descent Boosting

- ① **Initialization:** $m := 0$; Initialize additive predictor with offset value

$$\hat{\eta}^{[0]}(\cdot) \equiv \operatorname{argmin}_c \frac{1}{n} \sum_{i=1}^n \rho(y_i, c)$$

- ② **Negative gradient:** $m := m + 1$;

$$u_i^{[m]} = - \left. \frac{\partial \rho(y_i, \eta)}{\partial \eta} \right|_{\eta = \hat{\eta}^{[m-1]}(\mathbf{x}_i)}, \quad i = 1, \dots, n$$

- ③ **Estimation:** Fit the negative gradient vector $\mathbf{u}^{[m]} = (u_1^{[m]}, \dots, u_n^{[m]})$ to $\mathbf{x}_1, \dots, \mathbf{x}_n$ by real-valued base-learners **separately**:

$$(\mathbf{x}_i, u_i^{[m]})_{i=1}^n \xrightarrow{\text{base-learner}} \hat{g}_j^{[m]}(\cdot), \quad j = 1, \dots, J$$

- 4 **Selection:** Chose best fitting base-learner g_{j^*} with respect to some criterion (classically g_{j^*} that minimizes RSS):

$$j^* = \operatorname{argmin}_{1 \leq j \leq p} \sum_{i=1}^n (u_i^{[m]} - g_j(x_{ij}, \hat{\beta}_j^{[m]}))^2.$$

- 5 **Update:** Compute the update for the additive predictor

$$\hat{\eta}^{[m]}(\cdot) = \hat{\eta}^{[m-1]}(\cdot) + \nu \cdot \hat{g}_{j^*}^{[m]}(\cdot)$$

with step-length factor $0 < \nu \leq 1$.

- 6 **Stopping rule:** Continue iterating steps (2) to (5) until $m = m_{\text{stop}}$ for a given stopping iteration m_{stop} .

We stated that we use . . .

. . . component-wise boosting as a means of **estimation** and **variable selection** combined with **model choice**.

But how?

. . . is achieved by

- selection of base-learner, i.e., **component-wise boosting**

and

- **early stopping**,
i.e., choose $\hat{m}_{\text{stop,opt}}$ via cross validation, out-of-bag sample, . . .

We stated that we use . . .

. . . component-wise boosting as a means of **estimation** and **variable selection** combined with **model choice**.

Variable Selection and Model Choice

. . . is achieved by

- selection of base-learner, i.e., **component-wise boosting**

and

- **early stopping**,
i.e., choose $\hat{m}_{\text{stop,opt}}$ via cross validation, out-of-bag sample, . . .

Biased Selection of Categorical Covariates

- Problem: Covariate with many categories has higher flexibility
($df = n_{\text{cat}} - 1$)
- Thus: Preferred selection

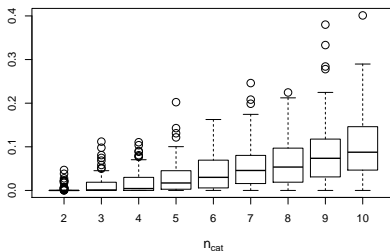
A measure for **selection bias**: $MSE = \frac{1}{p} \sum_{i=1}^p (\hat{\beta}_i - \beta_i)^2$

Biased Selection of Categorical Covariates

- Problem: Covariate with many categories has higher flexibility (df = $n_{\text{cat}} - 1$)
- Thus: Preferred selection

A measure for selection bias: $\text{MSE} = \frac{1}{p} \sum_{i=1}^p (\hat{\beta}_i - \beta_i)^2$

$\text{MSE}_{\beta_{\text{cat}}}$ with increasing n_{cat}



A Solution – Ridge-Penalized Base-Learner

- Let \mathbf{Z} be the **dummy coded design matrix** of categorical covariate,
- $\mathbf{u}^{[m]}$ is the negative gradient in iteration m .
- Classical **OLS base-learner**:

$$\hat{\mathbf{u}}^{[m]} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{u}^{[m]}$$

- Replace by **ridge-penalized base-learner**: (Hoerl & Kennard, 1970)

$$\hat{\mathbf{u}}^{[m]} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{I})^{-1} \mathbf{Z}^\top \mathbf{u}^{[m]}$$

where $\mathbf{I} = \text{diag}(1, \dots, 1)$

- Ridge-penalty shrinks parameter estimates towards zero.
- Shrinkage parameter λ is chosen according to pre-specified degrees of freedom.

⇒ **Use ridge-penalized base-learner with 1 df** to make the base-learners **comparable (w.r.t. df)**.

Basic Simulation Setting

Basic Model:

$$\mathbf{y} = \mathbf{X}^\top \boldsymbol{\beta} + \mathbf{Z}^\top \boldsymbol{\gamma} + \varepsilon$$

with

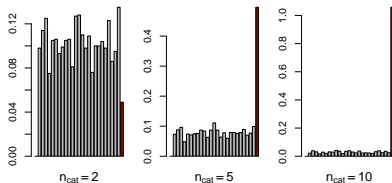
- Response vector \mathbf{y}
- Design matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{25})$, realizations from

$$X_1, \dots, X_{25} \stackrel{i.i.d.}{\sim} U[0, 1]$$

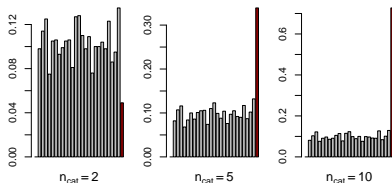
- Dummy coded design matrix \mathbf{Z} for z_1 with realizations from a discrete uniform distribution on $\{1, \dots, n_{\text{cat}}\}$
- Varying number of categories $n_{\text{cat}} \in \{2, \dots, 10\}$
- $\boldsymbol{\beta} = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$
- $\boldsymbol{\gamma}$ depends on the setting

- $\varepsilon \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$
- σ^2 such that the fraction of explained variance is $R^2 \approx 0.3$ (or $R^2 \approx 0.5$)
- $n = 150$ observations and $B = 100$ simulation replicates
- $\hat{m}_{\text{stop,opt}}$ determined based on an independent test sample of size $4n$ (i.e., 600 here)
- **X** enters the model centered!

Null Model Case $\beta = \mathbf{0}$ and $\gamma = \mathbf{0}$

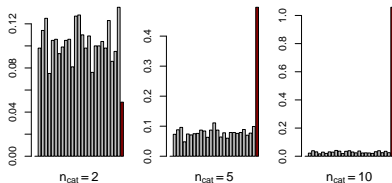
 $\hat{m}_{\text{stop,opt}}$


unpenalized model

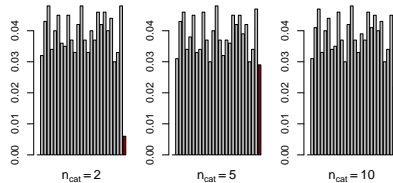
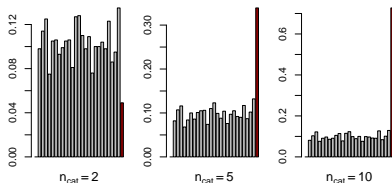
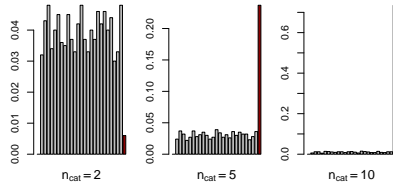


penalized model (i.e., model with ridge-penalized base-learner)

Null Model Case $\beta = \mathbf{0}$ and $\gamma = \mathbf{0}$

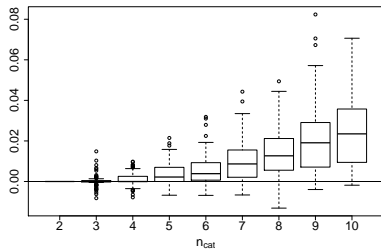
 $\hat{m}_{\text{stop,opt}}$


first step



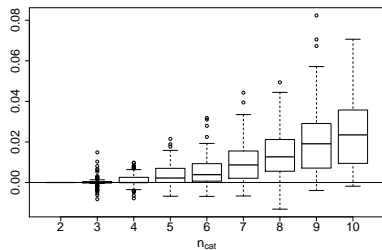
Power Case (1) $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$ and $\gamma = \mathbf{0}$

$\text{MSE}_{\text{unpenalized model}} - \text{MSE}_{\text{penalized model}}$

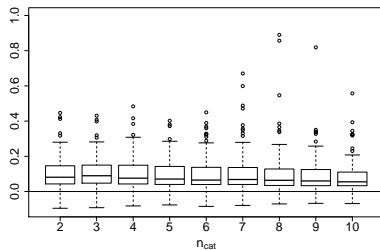


Power Case (1) $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$ and $\gamma = \mathbf{0}$

$\text{MSE}_{\text{unpenalized model}} - \text{MSE}_{\text{penalized model}}$



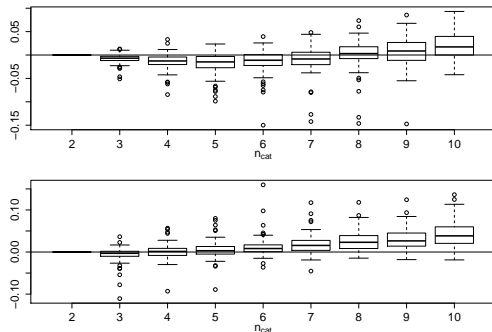
$\text{MSE}_{\text{linear model}} - \text{MSE}_{\text{penalized model}}$



Power Case (2) $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$, $\gamma_1 = \mathbf{0}$,

$$\gamma_2 = \left(\frac{2 \cdot 2}{n_{\text{cat}}}, \frac{2 \cdot 3}{n_{\text{cat}}}, \dots, \frac{2 \cdot n_{\text{cat}}}{n_{\text{cat}}} \right)^\top$$

$\text{MSE}_{\text{unpenalized model}} - \text{MSE}_{\text{penalized model}}$



MSE computed with

and without the influential,
categorical covariate
 Z_2

Increase of $\text{MSE}_{\text{penalized model}}$ due to shrinkage of $\hat{\gamma}_2$ **but** “remaining model improved”

Biased Selection of Smooth Effects

Degrees of freedom for linear effects ($df = 1$) and smooth effects ($df \gg 1$) are not comparable

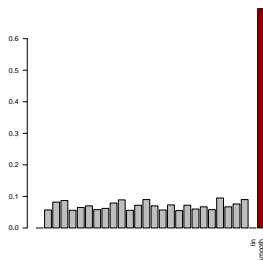
- Problem: We cannot make df of smooth effects arbitrary small, i.e., $df > 1$ ($\lambda \rightarrow \infty$) (for penalties of order $d \geq 2$)
- Hence: Polynomial of order $d - 1$ remains unpenalized

Biased Selection of Smooth Effects

Degrees of freedom for linear effects ($df = 1$) and smooth effects ($df \gg 1$) are not comparable

- Problem: We cannot make df of smooth effects arbitrary small, i.e., $df > 1$ ($\lambda \rightarrow \infty$) (for penalties of order $d \geq 2$)
- Hence: Polynomial of order $d - 1$ remains unpenalized

If we use more flexible base-learners (e.g., $df = 4$) the selection is biased



A measure for selection bias (with smooth effects)

- L_2 -norm of the deviation of the estimated (partial) function from the true function

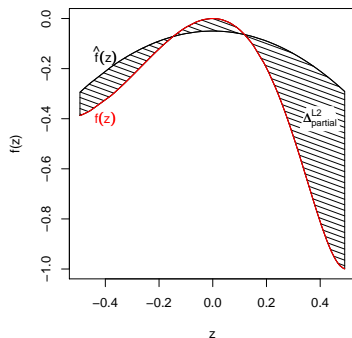
$$\Delta_{\text{partial},i}^{L_2} = \int_{\text{range}(\mathbf{x}_i)} [\hat{f}_i(\tilde{\mathbf{x}}) - f_i(\tilde{\mathbf{x}})]^2 d\tilde{\mathbf{x}}.$$

- Summary measure: mean L_2 deviation

$$\Delta^{L_2} = \frac{1}{p} \sum_{i=1}^p \Delta_{\text{partial},i}^{L_2},$$

where p is the number of covariates

- Very similar to the MSE



A Solution – P-Spline Decomposition

- For **model choice** we apply the decomposition

$$f_{\text{smooth}}(x) = \underbrace{\beta_0 + \beta_1 x}_{\text{unpenalized, parametric part}} + \underbrace{f_{\text{smooth,centered}}(x)}_{\text{deviation from polynomial}}$$

(based on Kneib, Hothorn, & Tutz, 2008)

- Add unpenalized part as separate, parametric base-learners
- Assign $df = 1$ to the centered effect (and add as P-spline base-learner)

Thus:

- Modeling components are **comparable** (w.r.t. df)
- Model choice between: linear effects and smooth effects. Further modeling alternatives, as varying coefficient terms, spatial effects, ... can be added analogously

Technical realization (see Fahrmeir, Kneib, & Lang, 2004):

decomposing the vector of regression coefficients β into $(\tilde{\beta}_{\text{unpen}}, \tilde{\beta}_{\text{pen}})$ utilizing a spectral decomposition of the penalty matrix

A Solution – P-Spline Decomposition

- For **model choice** we apply the decomposition

$$f_{\text{smooth}}(x) = \underbrace{\beta_0 + \beta_1 x}_{\text{unpenalized, parametric part}} + \underbrace{f_{\text{smooth,centered}}(x)}_{\text{deviation from polynomial}}$$

(based on Kneib et al., 2008)

- Add unpenalized part as separate, parametric base-learners
- Assign $df = 1$ to the centered effect (and add as P-spline base-learner)

Thus:

- Modeling components are **comparable** (w.r.t. df)
- Model choice between: linear effects and smooth effects. Further modeling alternatives, as varying coefficient terms, spatial effects, ... can be added analogously

Technical realization (see Fahrmeir et al., 2004):

decomposing the vector of regression coefficients β into $(\tilde{\beta}_{\text{unpen}}, \tilde{\beta}_{\text{pen}})$ utilizing a spectral decomposition of the penalty matrix

Basic Simulation Setting

Basic Model:

$$\mathbf{y} = \mathbf{X}^\top \boldsymbol{\beta} + f_Z(z_1) + \varepsilon$$

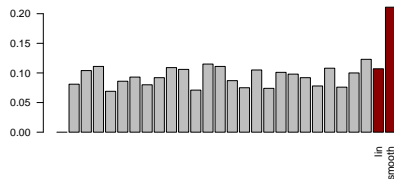
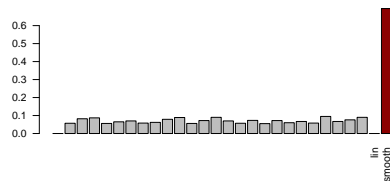
with

- Response vector \mathbf{y}
- Design matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{25})$ (as before)
- $\boldsymbol{\beta} = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$
- z_1 i.i.d. realizations from $Z \sim U[0, 1]$
- $f_Z(\cdot)$ depends on the setting
- \mathbf{X} and z_1 enter the model centered!

- $\varepsilon \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$
- σ^2 such that the fraction of explained variance is $R^2 \approx 0.3$ (or $R^2 \approx 0.5$)
- $n = 150$ observations and $B = 100$ simulation replicates
- $\hat{m}_{\text{stop,opt}}$ determined based on an independent test sample of size $4n$

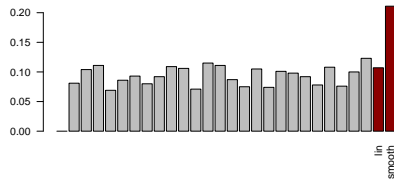
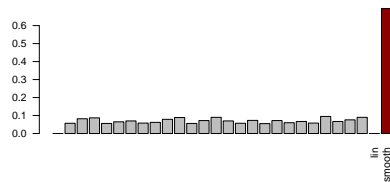
Null Model Case $\beta = \mathbf{0}$ and $f_Z(z_1) \equiv 0$

$\hat{m}_{\text{stop,opt}}$ (covariates centered)

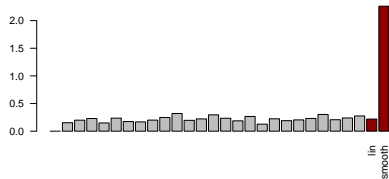
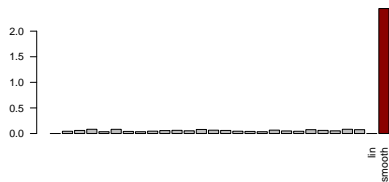


Null Model Case $\beta = \mathbf{0}$ and $f_Z(z_1) \equiv 0$

$\hat{m}_{\text{stop,opt}}$ (covariates centered)



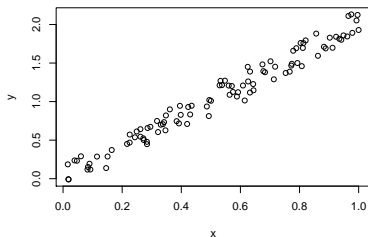
$\hat{m}_{\text{stop,opt}}$ (covariates not centered)



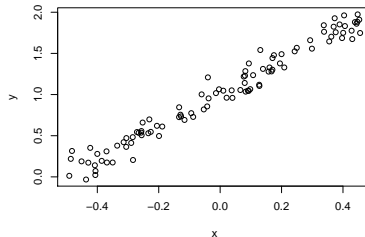
Importance of Centering

negative gradient and estimated base-learner

covariates not centered



covariates centered

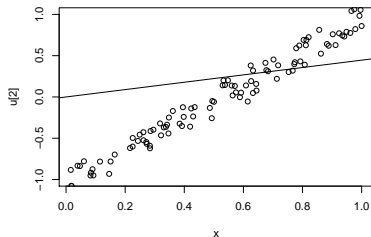


step
1

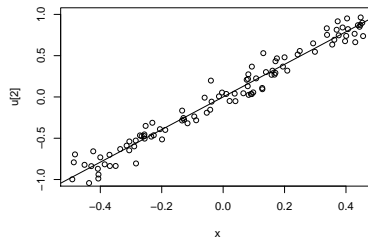
Importance of Centering

negative gradient and estimated base-learner

covariates not centered



covariates centered

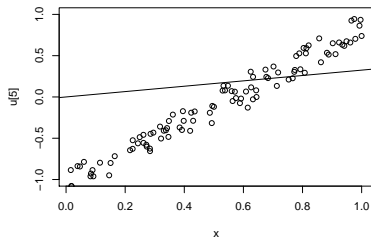


step
2

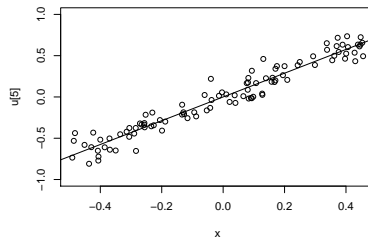
Importance of Centering

negative gradient and estimated base-learner

covariates not centered



covariates centered

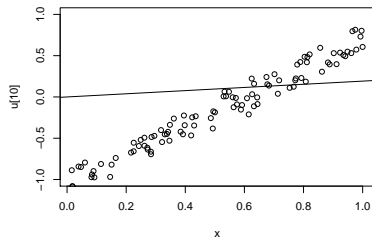


step
5

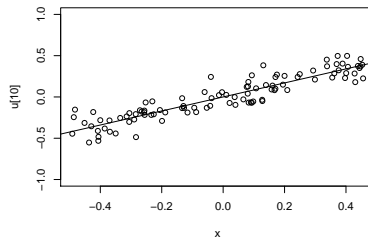
Importance of Centering

negative gradient and estimated base-learner

covariates not centered



covariates centered

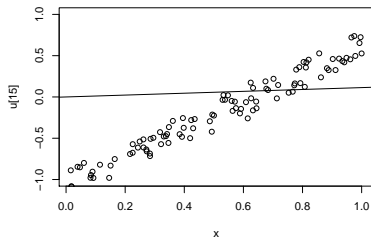


step
10

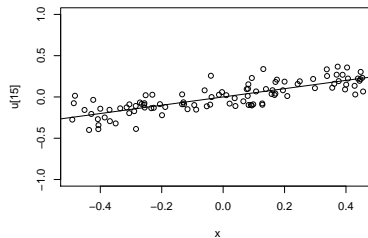
Importance of Centering

negative gradient and estimated base-learner

covariates not centered



covariates centered

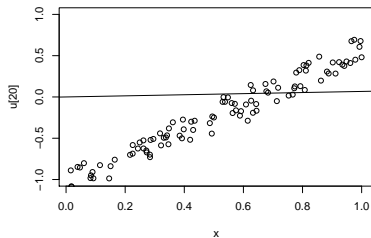


step
15

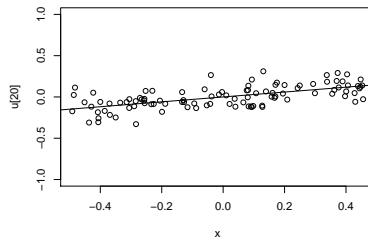
Importance of Centering

negative gradient and estimated base-learner

covariates not centered



covariates centered

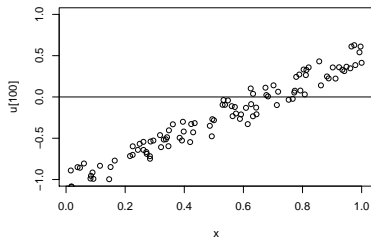


step
20

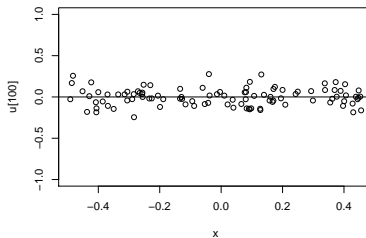
Importance of Centering

negative gradient and estimated base-learner

covariates not centered



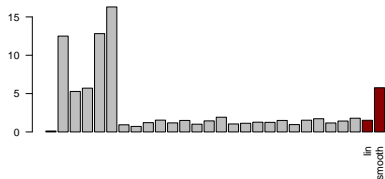
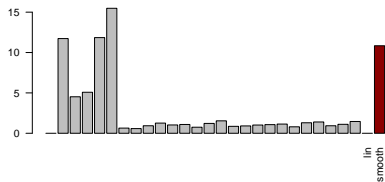
covariates centered



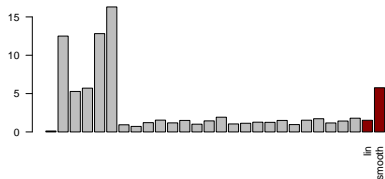
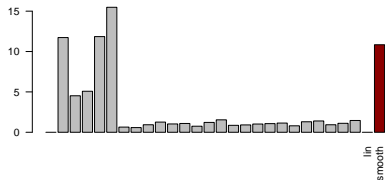
step
100

This happens because base-learner has [no intercept](#)

Power Case (1) z_1 non-influential, $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$, $f_Z(z_1) \equiv 0$

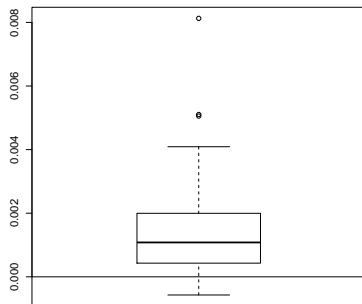
 $\hat{m}_{\text{stop,opt}}$


Power Case (1) z_1 non-influential, $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$, $f_Z(z_1) \equiv 0$

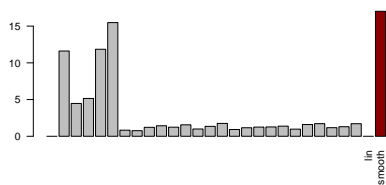
 $\hat{m}_{\text{stop,opt}}$


mean L_2 deviation:

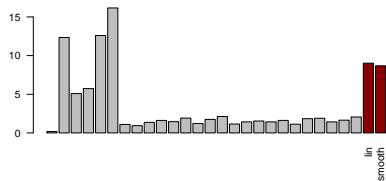
$$\Delta_{\text{model}}^{L_2} - \Delta_{\text{model with decomposition}}^{L_2}$$



Power Case (2) linear effect of z_1 , $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$, $f_Z(z_1) = 1.5z_1$

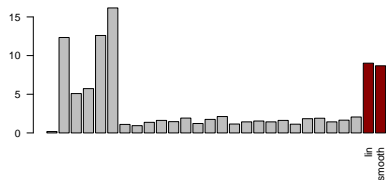
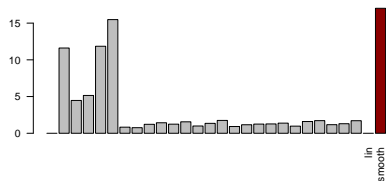
 $\hat{m}_{\text{stop,opt}}$


without decomposition
(linear + smooth with 4df)



with decomposition
(linear + smooth with 1df)

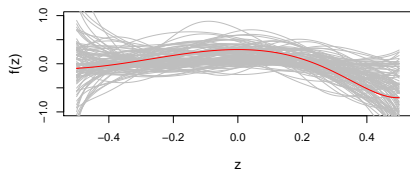
Power Case (2) linear effect of z_1 , $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$, $f_Z(z_1) = 1.5z_1$

 $\hat{m}_{\text{stop,opt}}$


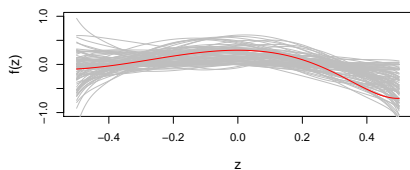
Power Case (3) smooth effect of z_1 , $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$,

$$f_Z(z_1) = \sin(-(2z_1)^2 - 0.6(2z_1)^3)$$

partial estimates of f_Z



without decomposition
(linear + smooth with 4df)

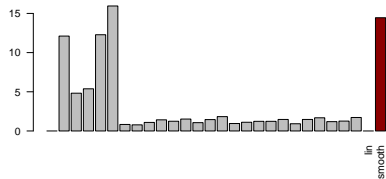


with decomposition
(linear + smooth with 1df)

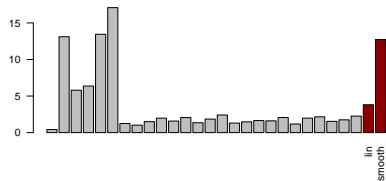
Power Case (3) ctd. smooth effect of z_1 , $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$,

$$f_Z(z_1) = \sin(-(2z_1)^2 - 0.6(2z_1)^3)$$

$\hat{m}_{\text{stop,opt}}$



without decomposition
(linear + smooth with 4df)

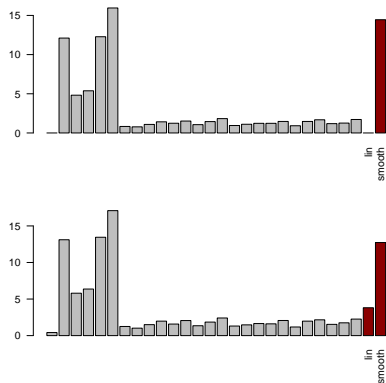


with decomposition
(linear + smooth with 1df)

Power Case (3) ctd. smooth effect of z_1 , $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$,

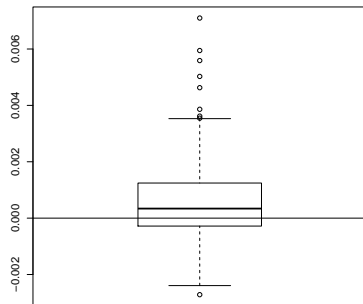
$$f_Z(z_1) = \sin(-(2z_1)^2 - 0.6(2z_1)^3)$$

$\hat{m}_{\text{stop,opt}}$



mean L_2 deviation:

$$\Delta_{\text{model}}^{L_2} - \Delta_{\text{model with decomposition}}^{L_2}$$



Forest Health Data - Results

We use a model where the P-spline decomposition and ridge-penalized base-learners are used to model the defoliation indicator.

Early stopping via cross validation leads to a model with:

Parametric effects for fertilisation, thickness of humus layer, and base saturation

Nonparametric effect(s) for canopy density (almost no effect: ph-value and soil depth)

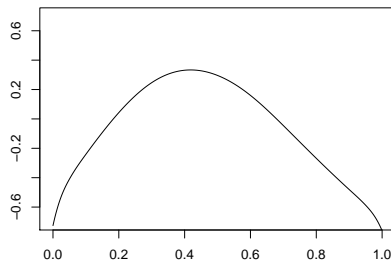
Spatial effect + unstructured random effect (with a clear domination of the latter)

Interaction effect between age and calendar time

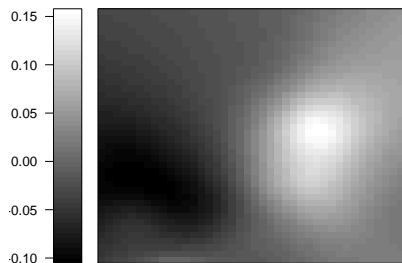
Not selected: type of stand, inclination of slope and elevation above sea level

Forest Health Data - Results (ctd.)

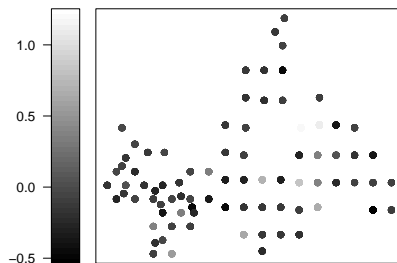
canopy density



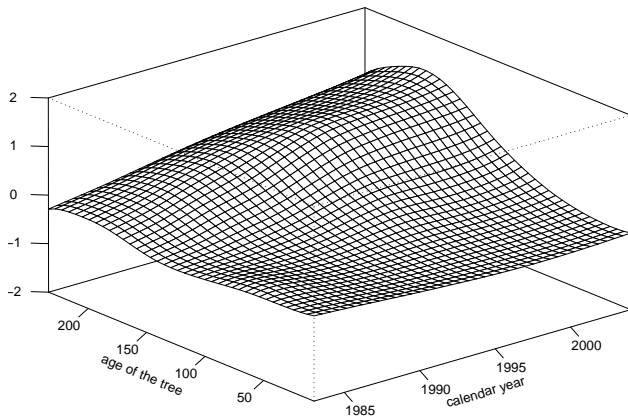
spatial effect



random effect



Forest Health Data - Results (ctd.)



Messages “To Go”

- One can fit a wide range of models by boosting:
 - different loss functions (LMs, GLMs, ...)
 - different base-learners (linear predictors, additive predictors, ...)
- If one uses linear or smooth base-learners (i.e., no tree base-learners) one gets interpretable models.

Problems: Variable Selection Bias (1) & Model Selection Bias (2)

- Improvement in both cases: Add (continuous) covariates centered.

Further improvements by using comparable (w.r.t. df) base-learners:

- **Solution for (1):** Add categorical covariates with ridge penalized base-learners with 1 df.
- **Solution for (2):** Add smooth effects after P-spline decomposition with 1 df.

R-package **mboost** available to fit all the models covered in this talk (Hothorn, Bühlmann, Kneib, Schmid, & Hofner, 2009)

Literature

- Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*, 89–121.
- Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica*, *14*, 731–761.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.
- Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2009). *mboost: Model-based boosting*. (R package version 1.1-1)
- Kneib, T., Hothorn, T., & Tutz, G. (2008). Variable selection and model choice in geoaddivitive regression models. *Biometrics*. ((accepted))

Find out more: <http://benjaminhofner.de/>