

# Biased Model Selection: Possible Solutions for Boosting

Benjamin Hofner <sup>1</sup>

Institut für Medizininformatik, Biometrie und Epidemiologie (IMBE)  
Friedrich-Alexander-Universität Erlangen-Nürnberg

joint work with  
Torsten Hothorn, Thomas Kneib and Matthias Schmid

Statistical Computing  
41st Workshop  
Reisensburg - 2009

---

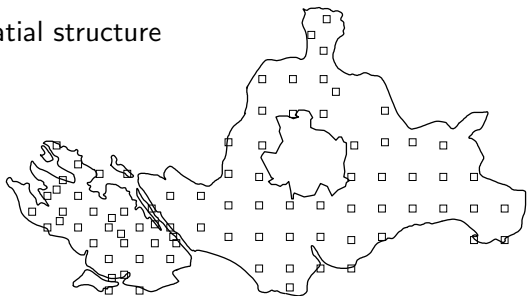
<sup>1</sup>[benjamin.hofner@imbe.med.uni-erlangen.de](mailto:benjamin.hofner@imbe.med.uni-erlangen.de)

# Aims

- Fitting models for (potentially) **high-dimensional data sets**
  - 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
  - **Response:** binary defoliation indicator at plot  $i$  in year  $t$ : ( $y_{it} = 1$  defoliation above 25%)
  - Large data set ( $n = 1793$ )
- ⇒ 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 a Solution)

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

# 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**:  $f_j(\mathbf{x}_i) = f_{j,\text{smooth}}(\tilde{x}_i)$   
(estimated using P-splines)

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).

## Boosting in a Nutshell

- Model fitting by: Minimizing the empirical risk
- Achieved by: Repeatedly fitting base-learners to the **negative gradient** of a pre-specified **loss function**  $\rho$  e.g.,
  - squared error loss**  $\rho(y, f(\mathbf{x})) = (y - \mathbf{x}^\top \beta)^2$  for linear regression problems
  - negative log-likelihood** for GLMs

For model fitting with intrinsic 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

## Boosting in a Nutshell

- Model fitting by: Minimizing the empirical risk
- Achieved by: Repeatedly fitting base-learners to the **negative gradient** of a pre-specified **loss function**  $\rho$  e.g.,
  - **squared error loss**  $\rho(y, f(\mathbf{x})) = (y - \mathbf{x}^\top \beta)^2$  for linear regression problems
  - **negative log-likelihood** for GLMs

For model fitting with intrinsic 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)  $m := 0$ ; Initialize additive predictor with offset value  $\hat{\eta}^{[0]}$
- (2)  $m := m + 1$ ; Compute negative gradient evaluated at the estimates of the previous iteration:  $u_i^{[m]} = - \left. \frac{\partial \rho(y_i, \eta)}{\partial \eta} \right|_{\eta = \hat{\eta}^{[m-1]}(\mathbf{x}_i)}$ ,  $i = 1, \dots, n$
- (3) 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  $g_j(\cdot)$  **separately**.
- (4) **Choose best fitting base-learner**  $g_{j^*}$  that minimizes RSS

$$j^* = \operatorname{argmin}_{1 \leq j \leq p} \sum_{i=1}^n \left( u_i^{[m]} - \hat{g}_{j^*}^{[m]}(\mathbf{x}_{ij^*}) \right)^2.$$

- (5) 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) Iterate steps (2) to (5) until  $m = m_{\text{stop}}$  for a given stopping iteration  $m_{\text{stop}}$ .

## Remarks

- Use linear, smooth (P-splines) and categorical base-learners:  
⇒ **the resulting model is a structured additive model!**
- **Major tuning parameter:**  $m_{\text{stop}}$   
(choose  $\hat{m}_{\text{stop,opt}}$  that minimizes empirical risk on “new data” via cross validation, bootstrap, ...)
- Step-length  $\nu$  is “no real tuning parameter” but governs how fast the algorithm converges (as long as it is small enough)
- If we use the **L2 loss** (“linear regression case”), the negative gradient reduces to least squares residuals as in a LM.  
⇒ **Boosting can be regarded as refitting residuals.**
- **Variable selection / model choice** is achieved by the selection of the (best-fitting) base-learner in each step and early stopping.

## Biased Selection of Categorical Covariates

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

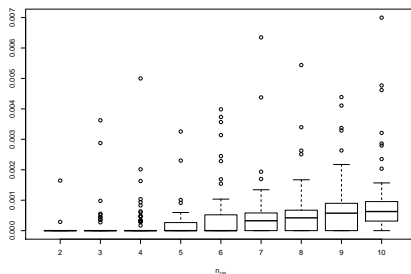
A measure for **selection bias**:  $MSE = \frac{1}{p} \sum_{i=1}^p (\hat{\beta}_i - \beta_i)^2$

# 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  $n_{\text{cat}}$



## A Solution – Ridge-Penalized Base-Learner

- Replace OLS base-learner

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

with ridge-penalized base-learner (Hoerl & Kennard, 1970)

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

where

- $\mathbf{Z}$  is the dummy coded design matrix of categorical covariate,
  - $\mathbf{u}^{[m]}$  is the negative gradient in iteration  $m$  and
  - $\mathbf{I} = \text{diag}(1, \dots, 1)$ .
- Ridge-penalty shrinks parameter estimates towards zero.
  - Shrinkage parameter  $\lambda$  is chosen according to pre-specified df.
- ⇒ Use ridge-penalized base-learner with 1 df to make the base-learners comparable (w.r.t. df).

Results:

- ✓ Empirical evaluation showed reduction of selection bias.

## A Solution – Ridge-Penalized Base-Learner

- Replace OLS base-learner

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

with 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

- $\mathbf{Z}$  is the dummy coded design matrix of categorical covariate,
  - $\mathbf{u}^{[m]}$  is the negative gradient in iteration  $m$  and
  - $\mathbf{I} = \text{diag}(1, \dots, 1)$ .
- Ridge-penalty shrinks parameter estimates towards zero.
  - Shrinkage parameter  $\lambda$  is chosen according to pre-specified df.
- ⇒ Use ridge-penalized base-learner with 1 df to make the base-learners comparable (w.r.t. df).

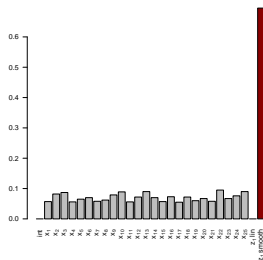
### Results:

- ✓ Empirical evaluation showed reduction of selection bias.

## Biased Selection of Smooth Effects

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

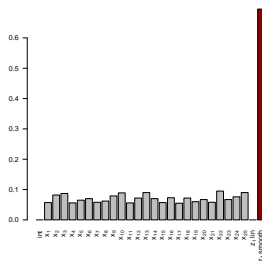
If we use more flexible base-learners (e.g.,  $df = 4$ ) the selection is biased



## Biased Selection of Smooth Effects

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

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



### Problem:

- We cannot make  $df$  of smooth effects arbitrary small ( $\lambda \rightarrow \infty \Rightarrow df > 1$ )

$\Rightarrow$  Linear effects remain unpenalized (in our case)

## 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, 2009)

- Add unpenalized part as **linear base-learner**
- Add centered effect  $f_{\text{smooth,centered}}(x)$  as **P-spline base-learner with 1 df**

**Thus:** Linear and smooth components are **comparable (w.r.t. df)**

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

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

## A measure for selection bias (with smooth effects)

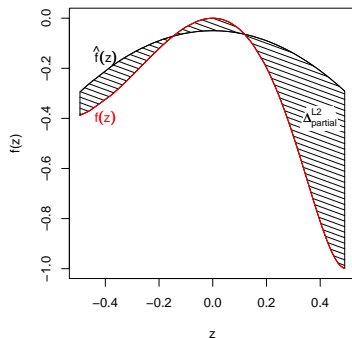
- $L_2$ -norm of the **deviation of the estimated (partial) function from the true function**

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

- Summary measure:  
mean  $L_2$  deviation from the true model

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

where  $p$  is the number of covariates



# Basic Simulation Setting

## Basic Model:

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

with

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

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

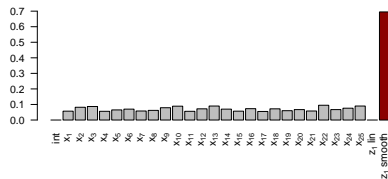
and  $z_1$  i.i.d. realizations from

$$Z \sim U[0, 1]$$

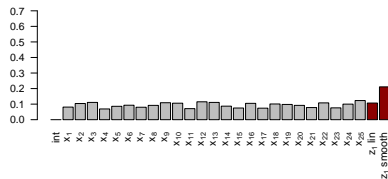
- $\boldsymbol{\beta} = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$
- $f_Z(\cdot)$  depends on the setting
- $\mathbf{X}$  and  $z_1$  enter the model centered!
- $\varepsilon \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$  with  $\sigma^2$  such that  $R^2 \approx 0.3$
- $n = 150$  observations and  $B = 100$  simulation replicates
- $\hat{m}_{\text{stop,opt}}$  determined based on an independent test sample of size  $5n$

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

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

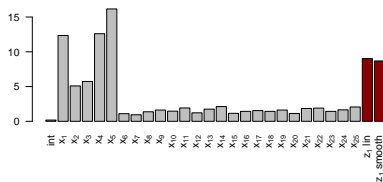
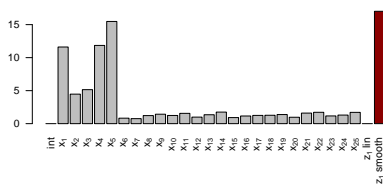


without decomposition  
(linear + smooth with 4df)



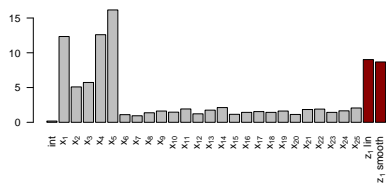
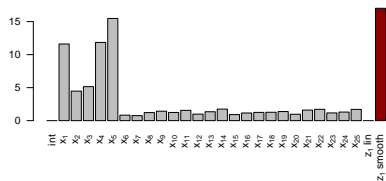
with decomposition  
(linear + smooth with 1df)

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

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


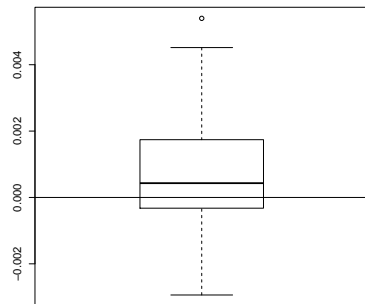
Power Case (1) linear effect of  $z_1$ ,  $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$ ,  $f_Z(z_1) = 1.5z_1$

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



mean  $L_2$  deviation:

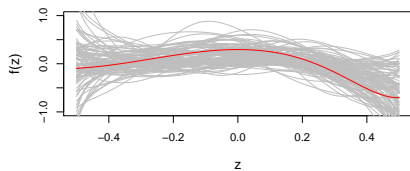
$\Delta_{\text{model}}^{L_2} - \Delta_{\text{model with decomposition}}^{L_2}$



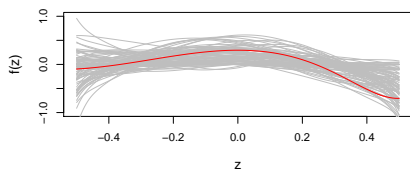
Power Case (2) smooth effect of  $z_1$ ,  $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$ ,

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

partial estimates of  $f_Z$



without decomposition  
(linear + smooth with 4df)

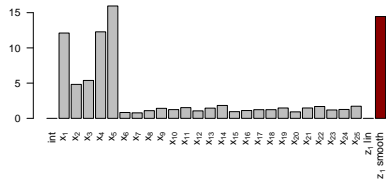


with decomposition  
(linear + smooth with 1df)

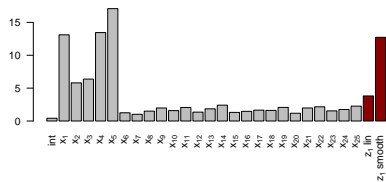
Power Case (2) ctd. smooth effect of  $z_1$ ,  $\beta = (-2, -1, 1, 2, 3, \mathbf{0}^\top)^\top$ ,

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

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



without decomposition  
(linear + smooth with 4df)

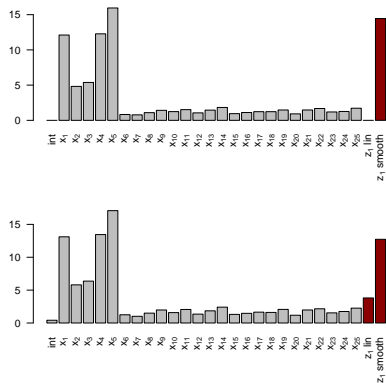


with decomposition  
(linear + smooth with 1df)

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

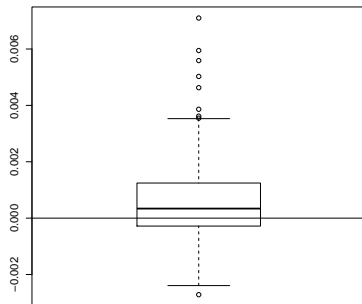
$$f_Z(z_1) = \sin(-(2z_1)^2 - 0.6(2z_1)^3)$$

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



mean  $L_2$  deviation:

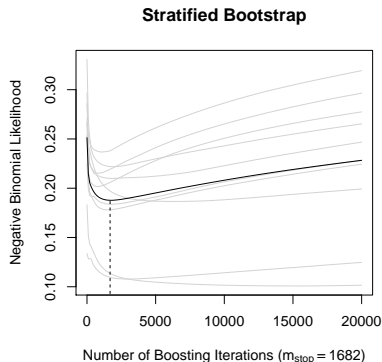
$$\Delta_{\text{model}}^{L_2} - \Delta_{\text{model with decomposition}}^{L_2}$$



## 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 stratified bootstrap (i.e., sampling from **plots** and **not from observations**)



## Forest Health Data - Results (ctd.)

Parametric effects for fertilisation and base saturation

Nonparametric effect for canopy density

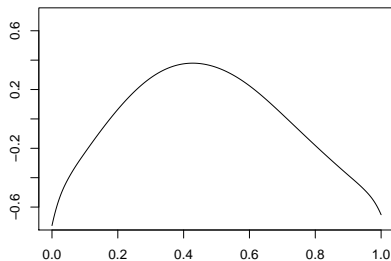
Spatial effect + unstructured random effect  
(with a clear domination of the latter)

Interaction effect between age and calendar time

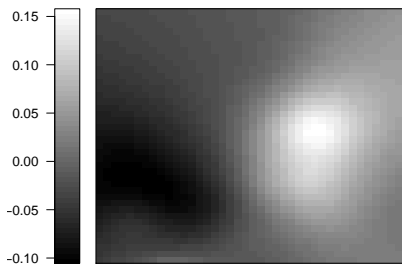
Not selected: thickness of humus layer, ph-value, soil depth, type of stand, inclination of slope, elevation above sea level

## Forest Health Data - Results (ctd.)

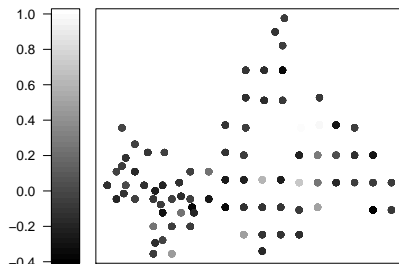
canopy density



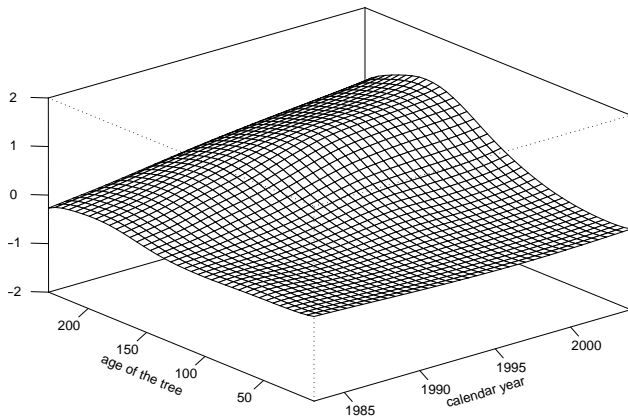
spatial effect



random effect



# Forest Health Data - Results (ctd.)



## Take-Away Messages

- One can fit a wide range of models by boosting: LMs, GLMs, survival models, GAMs, structured additive models, . . .
- Boosting results in interpretable models, if one uses linear or smooth base-learners (i.e., no tree base-learners).

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

**Improvements** by using comparable base-learners (w.r.t. df):

- **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

- 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. (2009). Variable selection and model choice in geoaddivitive regression models. *Biometrics*, 65, 626–634.

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