Benjamin Hofner, Torsten Hothorn,
Thomas Kneib & Matthias Schmid

A Framework for Unbiased Model Selection Based on Boosting

Technical Report Number 072, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

A Framework for Unbiased Model Selection Based on Boosting

Benjamin Hofner[*], Torsten Hothorn, Thomas Kneib, Matthias Schmid

Abstract

Variable selection and model choice are of major concern in many statistical applications, especially in high-dimensional regression models. Boosting is a convenient statistical method that combines model fitting with intrinsic model selection. We investigate the impact of base-learner specification on the performance of boosting as a model selection procedure. We show that variable selection may be biased if the covariates are of different nature. Important examples are models combining continuous and categorical covariates, especially if the number of categories is large. In this case, least squares base-learners offer increased flexibility for the categorical covariate and lead to a preference even if the categorical covariate is non-informative. Similar difficulties arise when comparing linear and nonlinear base-learners for a continuous covariate. The additional flexibility in the nonlinear base-learner again yields a preference of the more complex modeling alternative. We investigate these problems from a theoretical perspective and suggest a framework for unbiased model selection based on a general class of penalized least squares base-learners. Making all base-learners comparable in terms of their degrees of freedom strongly reduces the selection bias observed in naive boosting specifications. The importance of unbiased model selection is demonstrated in simulations and an application to forest health models.

Keywords: effective degrees of freedom, penalized least squares base-learner, penalized ordinal predictors, P-splines, ridge penalization, variable selection

1 Introduction

The methodological and computational advances in statistical regression modeling that we have seen during the last 15 years make it possible nowadays to model regression relationships in complex or high-dimensional structures that are hard to handle using the classical methods, such as GLMs with stepwise selection.

Especially ideas from computer science and machine learning have become popular in this respect. Perhaps the three most influential approaches are random forests (Breiman, 2001), support vector

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

machines (Vapnik, 1995), and boosting (Freund and Schapire, 1996). While random forest is a rather simple yet very powerful non-parametric approach to regression modeling, support vector machines and boosting might rather be seen as "meta-algorithms" that provide a rich framework to derive specialized solutions from.

The main focus of these methods, at least from a machine learning point of view, is prediction modeling, i.e., the construction of a superb oracle for unseen data. However, statisticians are more interested in inference about the unknown regression relationship. Random forests, support vector machines or boosting, however, do not necessarily provide measures statisticians are used to interpret. As a remedy, variable importance measures for random forests or tree-based gradient boosting are commonly used to pick "important" variables from the set of potentially many covariates.

Yet, there is ongoing discussion about the nature of such variable importance measures. Their theoretical foundations are still under debate (e.g., van der Laan, 2006) and some unintended behavior has been observed. The most problematic one is the so-called variable selection bias. Basically, a variable might receive a high variable importance not only because of their correlation with the response but also because of it's measurement scale. The problem has received a lot of attention in the regression tree community since the 1980s (Breiman *et al.*, 1984; Loh and Vanichsetakul, 1988; Loh, 2002; Kim and Loh, 2003; Hothorn *et al.*, 2006b) and, later on, was also observed and described for random forests (Strobl *et al.*, 2007).

Clearly, the problem comes from the fact that no well-defined statistical model is available that describes these methods in a probabilistic way. One way out of this dilemma was shown by the seminal papers of Friedman *et al.* (2000) and Bühlmann and Yu (2003) who interpreted functional gradient boosting as an optimization algorithm that can be modified in a way such that the resulting fit can be reformulated into a generalized additive model. Consequently, statisticians can interpret these models based on regression coefficients (linear model) or by looking at the partial contributions of each model component. In the meantime, boosting procedures for advanced model fitting have been introduced to a variety of fields, for example survival analysis (e.g. Hothorn *et al.*, 2006a; Schmid and Hothorn, 2008b) or spatial statistics (Kneib *et al.*, 2009).

The finding we are going to present in this paper is the fact that generalized additive models fitted using a component-wise functional gradient boosting algorithm are also, under circumstances described later, subject to variable selection bias. One instance of the problem is that a categorical covariate with a large number of levels gets selected more often compared to a covariate that has the same "importance" but is measured at less levels. We investigate the sources of variable selection bias in component-wise boosting theoretically. The results give insights into how to modify the algorithm to reduce the effect of variable selection bias. We finally study the effect empirically in artificial data generating processes and present a case-study on forest health where a complex spatial regression model not suffering from variable selection bias is fitted.

2 Component-Wise Boosting for Regression Models

Consider observations $(y_i, \boldsymbol{x}_i^\top), i = 1, \dots, n$, where y_i is the response variable and $\boldsymbol{x}_i^\top$ consists of possible predictors of different nature, such as categorical and continuous covariates. To model the dependence of the response on the predictor variables, we consider a structured regression model where $\mathbb{E}(y|\boldsymbol{x}) = h(\eta(\boldsymbol{x}))$ with (known) response function h and structured additive predictor $\eta(\boldsymbol{x})$ of the form

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{J} f_j(\boldsymbol{x}). \tag{1}$$

The functions $f_j(\cdot)$ are generic representations for modeling alternatives such as linear effects ($f_j(\boldsymbol{x}) = x\beta$, where x is one of the predictors), categorical effects ($f_j(\boldsymbol{x}) = \boldsymbol{z}^\top\boldsymbol{\beta}$, where $\boldsymbol{z}$ results from dummy-coding of a categorical covariate) and smooth effects ($f_j(\boldsymbol{x}) = f_{j,\text{smooth}}(x)$, where x is one of the predictors). Other modeling alternatives such as spatial and random effects can also be expressed in this framework, see Fahrmeir *et al.* (2004) for details. Generalized additive models (GAMs) as introduced by Hastie and Tibshirani (1986, 1990) appear as an important special case of (1). Having the model formulation at hand, two challenges arise: First, a method for model fitting in this flexible framework is needed. Second, the question which covariates should enter the model and how these covariates should be modeled needs to be answered. All issues can be addressed in one framework applying component-wise boosting.

Component-wise functional gradient descent boosting (see e.g., Bühlmann and Hothorn, 2007, for a detailed introduction) aims at minimizing the expected loss $\mathbb{E}(\rho(y, \eta))$ with respect to the structured predictor η, where $\rho(\cdot, \cdot)$ is a suitable loss function for the statistical model under consideration, such as the L_2-loss for "Gaussian" regression problems or the negative log-likelihood in more general cases. In practice, minimization of the expected loss is replaced by minimizing the empirical risk $n^{-1} \sum_{i=1}^{n} \rho(y_i, \eta_i)$ by component-wise boosting. After initialization of the function estimates $\hat{f}_j^{[0]}(\cdot) \equiv 0$ and the additive predictor $\hat{\eta}^{[0]}(\cdot) \equiv \operatorname{argmin}_c \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, c)$ the negative gradient of the loss function $\rho(y_i, \eta)$ is computed and evaluated at the predicted values of the previous iteration $\hat{\eta}^{[m-1]}(\boldsymbol{x}_i)$:

$$u_i^{[m]} = - \left. \frac{\partial \rho(y_i, \eta)}{\partial \eta} \right|_{\eta = \hat{\eta}^{[m-1]}(\boldsymbol{x}_i)}, \; i = 1, \dots, n. \tag{2}$$

We then relate the negative gradient vector $\boldsymbol{u}^{[m]} = (u_1^{[m]}, \dots, u_n^{[m]})'$ to subsets of the covariates using real-valued base-learners g_j, usually by least squares or penalized least squares estimation (Bühlmann and Yu, 2003). The base-learners correspond to the modeling alternatives as expressed by the generic functions f_j in the structured predictor (1), although each effect in (1) may be represented by more than one base-learner for example if model choice between competing modeling alternatives shall be implemented.

After evaluating all base-learner, we choose the best fitting g_{j^*}, i.e., the base-learner that minimizes the residual sum of squares (RSS)

$$j^* = \underset{1 \leq j \leq J}{\operatorname{argmin}} \sum_{i=1}^{n} (u_i^{[m]} - g_j(\boldsymbol{x}_i))^2 \tag{3}$$

and compute the update of the additive predictor $\hat{\eta}^{[m]}(\cdot) = \hat{\eta}^{[m-1]}(\cdot) + \nu \cdot \hat{g}_{j^*}^{[m]}(\cdot)$ and the function estimate $\hat{f}_{j^*}^{[m]}(\cdot) = \hat{f}_{j^*}^{[m-1]}(\cdot) + \nu \cdot \hat{g}_{j^*}(\cdot)$ while leaving all other function estimates f_j, $j \neq j^*$ unchanged. In each update step, only a fraction $0 < \nu \leq 1$ of the fitted values are added, which can be seen as a step-length factor in the gradient descent approach. As we select only *one* modeling alternative in each boosting iteration, variable selection and model choice is achieved by stopping the boosting procedure after an appropriate number of iterations $\widehat{m}_{\text{stop,opt}}$.

Obviously, the base-learner selection in (3) is the crucial part for variable and model selection. For variable selection, one specifies one base-learner for each covariate, while model choice is incorporated by additionally specifying base-learners for different, competing modeling alternatives (Kneib *et al.*, 2009).

All base-learners $g_j(\boldsymbol{x})$ considered in this paper can be expressed as penalized linear models

$$g_j(\boldsymbol{x}) = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{K})^{-1} \boldsymbol{X}^\top \boldsymbol{u}, \tag{4}$$

where $\boldsymbol{X}$ is a suitable design matrix for $\boldsymbol{x}$, λ is the smoothing parameter and $\boldsymbol{K}$ is a suitable penalty matrix. The smoothing parameter λ governs the amount of penalization and unpenalized least squares base-learners appear as a special case with $\lambda = 0$.

In the case of a continuous covariate x, we consider penalized least squares base-learners based on P-splines as introduced by Eilers and Marx (1996) for nonparametric regression and converted to the boosting framework by Schmid and Hothorn (2008a, see there for details). While an unpenalized least squares base-learner might be the first choice for (dummy coded) categorical covariates, we consider the more general approach of univariate ridge regression (Hoerl and Kennard, 1970) with treatment contrasts to serve as base-learners. In the case of ordinal categorical covariates, one could again use a ridge penalty for the coefficients of the dummy coded design matrix if penalized estimation is desired. However, it is often the case that ordering of the covariate categories converts to a similar ordering of the corresponding effects and this additional information can be incorporated to enforce stable estimation. Therefore, we consider a ridge-type penalty for the *differences* of adjacent parameters that favors smooth coefficient sequences similar as for penalized splines. A slight difference arises from the fact that the effect of the reference category is restricted to zero, and we will use the restriction $\beta_1 = 0$ in the following. Hence, the penalty is given by

$\sum_{i=2}^{n_{\text{cat}}}(\beta_i - \beta_{i-1})^2$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n_{\text{cat}}})^\top$ is the vector of dummy coded effects. For more details we refer to Gertheiss and Tutz (2009).

3 (Un-) Biased Selection of Base-Learners

Using the component-wise boosting approach naturally leads to variable selection and model choice if we choose an appropriate stopping iteration $\widehat{m}_{\text{stop,opt}}$. However, the selection of base-learners in each iteration can be seriously biased if the competing base-learners have different degrees of flexibility. This bias is intuitively plausible if one tries to distinguish whether a covariate x has a linear or a smooth effect on y. In this case, the usual strategy would be to specify a linear base-learner $g_1(x) = \beta x$ and a smooth base-learner $g_2(x) = f_{\text{smooth}}(x)$ and to distinguish between the two based on the selection in the boosting algorithm. However, the smooth base-learner offers much more flexibility and typically incorporates a linear effect for x as a special case. Hence, we can expect that boosting (almost) always prefers the smooth base-learner over the linear base-learner, regardless of the nature of the true effect. A similar selection bias can be expected when performing variable selection between competing categorical covariates with different numbers of categories. The covariate with more categories offers greater flexibility and thus is preferred in general when using unpenalized least squares base-learners.

In the following, we will theoretically investigate the presence of selection bias in the selection of base-learners for the special case of L_2-boosting based on the L_2-loss in the null model, i.e., when the response y is independent of the covariates.

Theorem 3.1 *Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be categorical covariates with M_1 and M_2 categories and design matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Let $\boldsymbol{u}$ be the $n \times 1$ negative gradient vector arising in the first step of the boosting algorithm for a response variable $\boldsymbol{y}$ of i.i.d. normally distributed random variables with variance σ^2 that is independent of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, i.e. $\boldsymbol{u}$ is simply the centered response variable. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ denote the effect estimates resulting from unpenalized least squares base-learners and* *define the difference of the residual sums of squares as* $\boldsymbol{\Delta} = (\boldsymbol{u} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{u} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1) - (\boldsymbol{u} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2)^\top(\boldsymbol{u} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2)$*. Then we have*

$$\mathbb{E}(\boldsymbol{\Delta}) = \sigma^2(M_2 - M_1), \tag{5}$$

i.e., $\mathbb{E}(\boldsymbol{\Delta}) = 0$ *if and only if* $M_1 = M_2$*.*

The proof can be found in Appendix A. Theorem 3.1 can be interpreted such that the expected difference of the RSS is greater than zero if the number of additional categories of $\boldsymbol{x}_2$, i.e. $M_2 - M_1$, is greater than zero, which reflects that a selection bias in favor of $\boldsymbol{x}_2$ is present. To overcome this problem, the base-learners should be made comparable with respect to their flexibility even if the number of categories is different. A specific possibility to achieve this is presented in the following theorem.

Theorem 3.2 *Assume that the assumptions from Theorem 3.1 hold. Furthermore, we replace the categorical base-learners with ridge penalized base-learners, where the penalty matrices* $\boldsymbol{K}_1$ *and* $\boldsymbol{K}_2$ *are identity matrices (of appropriate dimensions) and* λ_1 *and* λ_2 *are the corresponding smoothing parameters. Let* $\boldsymbol{S}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1 + \lambda_1\boldsymbol{K}_1)^{-1}\boldsymbol{X}_1^\top$ *be the smoother matrix of* $\boldsymbol{x}_1$ *and* $\boldsymbol{S}_2$ *be defined accordingly. Then*

$$\mathbb{E}(\boldsymbol{\Delta}) = 0 \Leftrightarrow \operatorname{tr}\left(2\boldsymbol{S}_1 - \boldsymbol{S}_1^\top\boldsymbol{S}_1\right) = \operatorname{tr}\left(2\boldsymbol{S}_2 - \boldsymbol{S}_2^\top\boldsymbol{S}_2\right), \tag{6}$$

where $\boldsymbol{\Delta}$ *is the difference in RSS resulting from the penalized least squares fits.*

The proof is again given in Appendix A. From Theorem 3.2 one can deduct that the degrees of freedom

$$\mathrm{df} := \operatorname{tr}\left(2\boldsymbol{S} - \boldsymbol{S}^\top\boldsymbol{S}\right) \tag{7}$$

should be comparable for the two competing base-learners in order to overcome the selection bias. Note that the degrees of freedom resulting from Theorem 3.2 are different from the standard definition in the smoothing literature given by $\widetilde{\mathrm{df}} := \operatorname{tr}(\boldsymbol{S})$. However, df is an alternative definition for the degrees of freedom in penalized models that is the preferred choice if one compares two models with respect to the RSS as stated by Buja *et al.* (1989) and confirmed by Theorem 3.2. In Web Supplement B, we show that a similar selection bias occurs when trying to distinguish between linear and smooth modeling alternatives based on penalized least squares base-learners and that the selection bias can be avoided by making the degrees of freedom df comparable. This can be seen as an improved version of the model choice scheme proposed in Kneib *et al.* (2009) who used $\widetilde{\mathrm{df}}$ instead of df. Following these lines, one should specify equal df for *all* base-learners if unbiased model choice and variable selection is the goal. The natural choice for this common degrees of freedom is one single free parameter as it appears for a simple least squares base-learner of one single continuous covariate. This can easily be achieved for categorical covariates by setting the smoothing parameter to an appropriate value (see below). Note that we do not include an intercept in the base-learners but specify a separate base-learner for the intercept.

However, for P-splines we cannot make df arbitrary small even with λ approaching infinity since a polynomial of order $d-1$ remains unpenalized by a d-th order difference penalty (Eilers and Marx, 1996). As we usually apply second order differences, a linear effect (with intercept) remains unpenalized and thus df ≥ 2 for all λ. To be able to specify a base-learner with df $= 1$ a reparameterization as described in Kneib *et al.* (2009) is needed, where the smooth base-learner is decomposed into parametric parts for the unpenalized polynomial and a smooth deviation from this polynomial

$$g_j(x) = \beta_{0,j} + \beta_{1,j}x + \ldots + \beta_{d-1,j}x^{d-1} + g_{\text{centered}}(x), \tag{8}$$

where only $g_{\text{centered}}(x)$ is modeled using a P-spline base-learner. Now, we can specify separate base-learners for each parametric effect and a base-learner with one degree of freedom for the smooth deviation from the polynomial. For more details on the technical realization and further implications of the decomposition we refer to Kneib *et al.* (2009).

As mentioned above, we specify the smoothness of all base-learners via the degrees of freedom. We use an initial value df_{init} for each penalized base-learner and solve $\mathrm{tr}\left(2\boldsymbol{S} - \boldsymbol{S}^\top\boldsymbol{S}\right) = df_{\text{init}}$ for

λ. The following lemma provides a convenient, numerically efficient way to compute the degrees of freedom and therefore to determine the corresponding λ.

Lemma 3.3 *Let $\boldsymbol{S} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}\boldsymbol{X}^\top$ be the smoother matrix of $\boldsymbol{x}$ with symmetric penalty matrix $\boldsymbol{K}$. Let $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{R}^\top\boldsymbol{R}$ be the Cholesky decomposition of the cross product of the design matrix. Then, the degrees of freedom* $\mathrm{df}(\lambda) = \mathrm{tr}\left(2\boldsymbol{S} - \boldsymbol{S}^\top\boldsymbol{S}\right)$ *are equal to*

$$\mathrm{df}(\lambda) = 2\sum_{j=1}^{M}\frac{1}{1+\lambda d_j} - \sum_{j=1}^{M}\frac{1}{(1+\lambda d_j)^2} \tag{9}$$

where $d_j \geq 0$ are the singular values of $\boldsymbol{R}^{-\top}\boldsymbol{K}\boldsymbol{R}^{-1}$.

The proof of Lemma 3.3 can be derived using the Demmler-Reinsch orthogonalization (cf., App. B.1.1 Ruppert *et al.*, 2003, with proof). As Lemma 3.3 only requires the penalty matrix $\boldsymbol{K}$ to be symmetric, we can use (9) to compute the degrees of freedom for all base-learners proposed in this paper.

4 Variable Selection Bias under Test

4.1 Biased Selection of Categorical Covariates

To empirically evaluate the bias introduced by categorical covariates with potentially many categories, we examine two situations: the null case, where *none* of the covariates has an influence on the response and a set of power cases, where a subset of the covariates influences the response. In the null case, the response is simply i.i.d. normally distributed, $\boldsymbol{y}_i \overset{i.i.d.}{\sim} N(0,1)$ but we fit a model with structure

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon} \tag{10}$$

where the $n \times p$ matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_p)$ is formed of continuous covariates, $\boldsymbol{Z}_1$ is a dummy coded design matrix for the categorical covariate z_1 and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_1$ are the corresponding parameter vectors. The $p = 25$ continuous covariates were sampled as realizations $X_1, \dots, X_p \overset{i.i.d.}{\sim} U[0,1]$ and the categorical covariate z_1, with varying numbers of categories $n_{\text{cat}} \in \{2, \dots, 10\}$, was sampled from a discrete uniform distribution on $\{1, \dots, n_{\text{cat}}\}$.

In the first power case, the response depends on five continuous covariates, the remaining 20 continuous covariates and the categorical covariate have no influence on y. The effects used to generate the response y can be found in Table 1 (upper part). Again, we fit a model with structure (10).

In the second power case setting, we add a second, informative categorical covariate z_2 with the same numbers of categories as z_1 and sampled i.i.d. from a discrete uniform distribution as used for z_1. The response now depends on five continuous covariates and on the categorical covariate z_2. The model is fitted according to the structure

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{Z}_2\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon} \tag{11}$$

where $\boldsymbol{Z}_2$ and $\boldsymbol{\gamma}_2$ are design matrix and coefficients vector for z_2. The effects of the categories for z_2 do not exceed two and hence are comparable to the other effects in (10). Table 1 summarizes all simulation settings.

For both the null case and the power cases, the sample size was set to $n = 150$ and the error terms are i.i.d. samples from $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with σ^2 chosen such that the fraction of explained variance is either $R^2 \approx 0.3$ or $R^2 \approx 0.5$ (results for the second case are omitted in the following because they are qualitatively the same as for $R^2 \approx 0.3$). In the power cases, we simulated $B = 100$ data sets while $B = 1000$ simulation replicates were considered in the null case.

All models were fitted using the `gamboost` function from the R package **mboost** and one separate base-learner was specified for each model component. All continuous covariates were standardized since we use base-learners without intercept (and specify an additional base-learner for the intercept). Without centering of the covariates, linear effects of base-learners without intercept would be forced through the origin (with no data lying there). Hence, convergence would be very slow or the algorithm would not converge to the "correct" solution even in very simple cases. For the base-learner of categorical covariates, we considered either unpenalized or penalized least squares base-learners. In the remainder of the paper we use the terms “unpenalized model” and “penalized model” to refer to models where the base-learner for the categorical effect is unpenalized and ridge penalized, respectively. The stopping iteration $\widehat{m}_{\text{stop,opt}}$ was determined based on an independent test sample of size 750.

To measure the *variable selection* bias we use the mean squared prediction error (MSE) of the coefficients

$$\text{MSE} = \frac{1}{\widetilde{p}} \sum_{i=1}^{\widetilde{p}} (\hat{\beta}_i - \beta_i)^2, \tag{12}$$

where $\widetilde{p}$ is the number of coefficients in the model including those for the categorical effect(s). Another important quantity is the selection frequency of the base-learners averaged over all simulation runs, which is a strong indicator for variable importance.

In the null model case, i.e., in the case where no covariate has an influence on the response, a sensible selection procedure should not prefer one base-learner over another but should randomly select any of the non-informative covariates. The selection rates in models with and without penalized base-learners can be found in Figure 1. Obviously, the selection frequency of the categorical covariate increases with increasing n_{cat} if no ridge penalty is applied. When applying the ridge penalty, the selection frequency of the categorical covariate becomes comparable to the selection frequency of the continuous covariates. Hence, we can conclude that using penalized categorical base-learners improves boosting algorithm in the null case w.r.t. the selection rates.

In the first power case, we again compare the performance of the boosting models with and without ridge penalization for categorical covariates. To correct for the bias, a ridge penalty is applied to the categorical covariate. Hence, the parameter estimates are shrunken such that the resulting dfs are all equal to one (independently of n_{cat}). Figure 2(a) indicates that the median selection rates are decreased when comparing the penalized and the unpenalized case and Figure 2(b) shows that the MSE is also decreased when using the penalized base-learner for z_1.

In the the second power case, an additional informative categorical covariate z_2 is included in the model. Figure 3(a) shows the difference of the MSE for the models with unpenalized and penalized base-learners for categorical covariates. As the boxes cover zero or sometimes are even located below, the "unpenalized model" seems to be better. However, the figures represent a mixture of two different, competing effects: The effect of the non-influential categorical covariate is shrunken towards zero with the penalized base-learner (decrease of MSE). At the same time, the effect of the influential covariate is also shrunken towards zero introducing an additional bias (increase of MSE). However, this is true for *any* penalization approach. Looking at Figure 3(b) one can see the difference of MSE where the influential, categorical covariate z_2 is excluded (for the calculation of the MSE but *not* for the estimation of the model). Here one clearly sees the superiority of the penalized approach. One can conclude that penalization has the advantage to reduce the selection bias for non-informative covariates and additionally shrinks the effects of influential covariates. If one deals with high-dimensional settings and variable selection and shrinkage are desired beforehand, the ridge penalty is exactly what one would like to apply.

To benchmark the model choice and variable selection scheme with ridge penalized base-learners we compared the resulting MSEs to the mean squared errors of a linear model with forward stepwise selection based on the AIC. In our settings, the boosting models are better on average than the stepwise models. In the first power case where the categorical covariate has no effect, the boosting model was better in more than 75% of the cases w.r.t. the mean squared prediction error. For the second power case with influential categorical covariate, boosting was superior to stepwise regression in more than 75% of the cases if we drop the informative categorical covariate for the computation of the MSE. If we compute the MSE with the informative categorical covariate, the shrinkage effect decreases the superiority of boosting a bit but still, boosting is superior to stepwise regression in the majority of the cases.

If categorical covariates are ordinal, one can use ordinal penalized base-learners instead of ridge penalized base-learners (see Sec. 2). To asses the properties of this penalty we used the same simulation setting as for unordered covariates (see Table 1).

To summarize the results (not shown here for sake of brevity), we can conclude that ordinal penalized base-learners show basically the same behavior as ridge penalized base-learners: In the null case, the penalized ordinal base-learners correct the selection bias such that the selection frequency of all base-learners is approximately equal. In the first power case, the MSE is improved in comparison to the unpenalized model. Figure 4(a) shows that both, ordinal penalized base-learner and ridge penalized base-learner are overall comparable in the first power case where the categorical covariate has no influence. In the second power case with an additional, informative covariate z_2, the penalized model shows again an improvement compared to the unpenalized model, but the ordinal penalty offers a further improvement over the ridge penalty, which does not exploit the ordinal structure of z_2. This can be seen in Figure 4(b), where we see another increase in the differences of the MSE. This is possibly due to a weaker penalization of the higher categories, which have a bigger effect (cf. Table 1): The ridge penalty shrinks all coefficients equally against zero, whereas the ordinal penalty just shrinks the increase with respect to the preceding category against zero. Hence, we can conclude that it is preferable to exploit the ordinal structure of the covariates if possible and only use ridge penalized base-learners if no ordinal structure can be assumed.

4.2 Biased Selection of Smooth Effects

To evaluate the preferred selection of smooth effects compared to linear effects we examined again the null case and a set of power cases. The data were generated similar to the categorical case (Sec. 4.1), where the categorical covariate was replaced by a continuous covariate. This leads to the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + f_z(z_1) + \boldsymbol{\varepsilon} \tag{13}$$

with the $n \times 1$ dimensional response vector $\boldsymbol{y}$, the $n \times p$ matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, and the corresponding parameter vector $\boldsymbol{\beta}$ of length p (see Table 1). The function $f_z(\cdot)$ can be of different natures, e.g., it can be a linear function, a smooth function or it can be a function that is equal to zero for all realizations $\boldsymbol{z}_1$ (cf. Table 1). The $p = 25$ continuous covariates were all sampled i.i.d. from $U[0,1]$, as well as the continuous covariate of interest z_1. The error term was sampled again from $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with σ^2 such that the fraction of explained variance is either $R^2 \approx 0.3$ or $R^2 \approx 0.5$. Results in the latter case are not reported here but are essentially the same as in the former case. To empirically evaluate the selection bias introduced by smooth terms compared to linear terms we simulated $B = 100$ data sets with $n = 150$ observations from model (13).

To measure the model *selection* bias that is introduced by competing linear and smooth base-learners we use the following L_2-norm

$$\Delta^{L_2}_{\text{partial},i} = \int_{\min(\boldsymbol{x}_i)}^{\max(\boldsymbol{x}_i)} [\hat{f}(\tilde{x}) - f(\tilde{x})]^2 d\tilde{x}. \tag{14}$$

Thus, we measure the deviation of the estimated partial function from the true function. For numerical evaluation, we predicted the model on a fine, equidistant grid and applied the trapezoidal rule to evaluate the integral. As a summary measure we use the mean L_2 deviation $\Delta^{L_2} = \tilde{p}^{-1} \sum_{i=1}^{\tilde{p}} \Delta^{L_2}_{\text{partial},i}$, where $\tilde{p}$ is the number of covariates. Thus, Δ^{L_2} can be seen as an analogon to the MSE (12) extended to smooth effects.

In the following paragraph, we exemplify the results for the second power case where $f_z(z_1) = 1.5 \cdot z_1$, i.e., the covariate of interest has a linear effect. We specify a linear and a smooth base-learner for z_1, once without applying the decomposition (8) and in the second scheme with decomposition (8). Figure 5(a) shows that the model with decomposition is almost always better (or comparable) to the model without decomposition with respect to the deviations of the partial fits, i.e. the fitted functions seem to be better in many cases. At the same time, we can observe a significant increase of the selection frequency of the linear base-learner for z_1 if the decomposition is applied. We can still observe a reasonable amount of selections for the smooth base-learner but one should note that the linear base-learner is selected more often than the smooth base-learner. Without the decomposition, the linear base-learner is never selected. Hence, the true nature of the underlying effect is missed in this case.

For the other simulation settings given in Table 1 (lower part; results not presented here) we observed that the P-spline decomposition (8), where all base-learners are specified with one degree of freedom, leads to an improved selection of modeling alternatives as well as improved models in terms of the mean L_2 deviation. In all simulations, we specified a linear and a smooth base-learner for z_1, once without applying the decomposition and once with the P-spline decomposition. In the null case, the model without decomposition showed a serious selection bias in favor of the smooth effect which vanishes if the decomposition is used. In the power case with non-influential z_1, the selection bias is again corrected by applying the decomposition and the mean L_2 deviation Δ^{L_2} is reduced, i.e. the model is improved. In the power case with smooth effect for z_1 both models are almost equally good (on average) regarding the mean L_2 deviation. Despite the fact that the model with P-spline decomposition reduces the selection of smooth effects as it makes them comparable to linear effects, the model without decomposition (and thus greater flexibility to model the smooth effect) is not better. Thus, the model with decomposition is clearly preferred. This is true for the given setting as well as for other models we investigated. Hence, we can conclude that using the model decomposition for P-splines leads to *overall improved models* and reduces the selection bias in favor of smooth effects.

5 Application: Forest Health Prediction

In our application, we consider models describing forest health status. The aim is to identify predictors of the health status of beeches, which is measured in terms of the degree of defoliation. The data originates from yearly visual forest health inventories carried out from 1983 to 2004 in a northern Bavarian forest district. The data consists of 83 plots of beeches within a 15 km × 10 km area with a total of $n = 1793$ observations. The response is a dichotomized version of the defoliation index indicating defoliation above 25%. Obviously, the data set combines a longitudinal and a spatial structure. An overview of the covariates is given in Table 2 (Web Supplement C).

Previous analyses (Kneib and Fahrmeir, 2006; Kneib *et al.*, 2009; Kneib and Fahrmeir, 2010) resulted in models that contained categorical covariates, as well as linear and smooth effects of continuous covariates. Additionally, a spatial effect and a random effect for the plots could be identified. When considering a structured additive regression model of comparable complexity in a naive boosting implementation, biased model selection of smooth model components as well as categorical covariates with several categories is likely to occur. In the following, we will apply the methodology developed in this paper to achieve unbiased variable selection and model choice for the forest health data. Since the outcome is binary, we minimize the negative binomial log-likelihood, i.e., we fit a structured logit model to the data.

We use linear, P-spline, ridge penalized and ordinal penalized base-learners to model the defoliation indicator. The spatial effect can be flexibly modeled using a tensor product of P-splines. Finally, a ridge penalized base-learner is assigned to the plot-specific (random) effects with a fixed smoothing parameter (see Kneib *et al.*, 2009, for details). All modeling alternatives considered for possible inclusion in the model can be specified as base-learners with 1 df and thus, the selection bias, as discussed in this paper, can be avoided. For more details on the candidate model we refer to Web Supplement C.

The optimal stopping iteration was estimated via stratified bootstrap, i.e., we randomly selected plots (with replacement) and not single observations, as the plots can be seen as the observational units. The resulting model included five covariates, the spatial information and the random intercept for the plots. Fertilization (represented as a binary indicator for the application of fertilization) was included in the model with a negative effect on defoliation ($\beta_{\text{fert}} = -0.76$), and age and calender time (both included as linear base-learners) had positive effects ($\beta_{\text{age}} = 0.016$ and $\beta_{\text{year}} = 0.068$). This means that the severity of defoliation increases each year if the other covariates (including the age of the trees in the plot) are kept fix. The effect estimates of base saturation, which was modeled using a penalized ordinal base-learner, and canopy density, which was included as a combination of a linear and a smooth base-learner, can be found in the upper part of Figure 6. The spatial effect was included in the model but is clearly dominated by the spatially unstructured, plot-specific effect (Figure 6). The remaining six covariates were not included in the final model. In summary, our boosting framework allows to fit a complex model comprising many different kinds of effects while obtaining results that are interpretable and biologically meaningful.

Finally, to benchmark our approach, we compared the bias corrected modeling approach with the uncorrected approach where each smooth base-learner and the random effect base-learner were added with 4 df and the categorical base-learners were added unpenalized. We used stratified 10-fold cross validation where the corrected as well as the uncorrected model were fitted on the learning sample. The optimal stopping iteration within each learning sample was estimated by stratified bootstrap separately for each model. Each of the 10 test samples was used to determine the out-of-bag risk, i.e., the negative log likelihood. We observed that the corrected model was superior to the uncorrected model in 80% of the cases with respect to the prediction error measured by the negative log likelihood. Thus, correcting for biased selection of base-learners resulted also in improved prediction accuracy.

6 Concluding Remarks

Component-wise boosting techniques offer the possibility to fit a wide range of models with intrinsic variable selection and model choice. To avoid selection bias of the base-learners, equal degrees of freedom need to be assigned to all base-learners. This can be achieved by using penalized least squares base-learners. We considered ridge penalized base-learners for categorical covariates, penalized base-learners with a ridge penalty applied to the differences of adjacent coefficients for ordinal covariates and penalized spline base-learners for smooth model terms. For the latter, an additional reparameterisation step has to be applied to differentiate between an unpenalized polynomial and the penalized deviation (Kneib *et al.*, 2009).

For all base-learners, degrees of freedom can be specified using $\mathrm{df} = \mathrm{tr}\left(2\boldsymbol{S} - \boldsymbol{S}^\top \boldsymbol{S}\right)$. This definition is tailored for the comparison of residual sums of squares (RSS) and also appeared naturally from our theoretical considerations about the selection bias. Furthermore, centering of covariates is highly important if base-learners without intercept (for each base-learner) are applied. It is a first, very important step to achieve unbiased model choice and variable selection using boosting. Specifying equal dfs for all modeling components to achieve unbiased model choice can be easily incorporated in all component-wise functional gradient descent boosting approaches.

In contrast, most of the literature dedicated to likelihood-based boosting (based on Fisher scoring) currently advocates to use one single smoothing parameter λ for the penalty (e.g., Tutz and Binder, 2006), or even to decrease the penalty for some covariates to prefer them in the selection step without a thorough theoretic reasoning, yielding a faster convergence towards the maximum (partial) likelihood estimates (Binder and Schumacher, 2008). Obviously, one could also define the smoothing parameters in this framework such that the resulting degrees of freedom are equal and thus obtain unbiased variable selection and model choice procedures based on likelihood-based boosting. However, one problem arises in this context: The degrees of freedom for a fixed smoothing parameter change over the subsequent boosting iterations (Hofner *et al.*, 2008), for example, due to changes in the working weights for GLMs.

Alternatively, one could think of altering the goodness-of-fit criterion that is used to determine the best fitting base-learner in each step instead of making the competing base-learners comparable by specifying equal dfs to achieve unbiased base-learner selection. Examples include penalized alternatives to the RSS such as AIC and BIC. In the context of likelihood-based boosting, Binder and Schumacher (2008) propose to use the *penalized* partial likelihood or the AIC as selection criterion but do not give empirical results for the selection frequencies of the base-learners. Another idea could be to use F-tests, which also account for the different degrees of freedom. However, our experience from simulation studies shows that such approaches are not working when comparing base-learners with very different degrees of freedom. Criteria such as AIC and BIC are composed of two parts: One that measures the fit via the likelihood and one term for the penalty. In the linear regression model, which applies here, we can write $\mathrm{AIC} = n/\log(\mathrm{RSS}/n) + 2\,\mathrm{df}$. In the course of the boosting procedure, the information still left in the data decreases sequentially and consequently the RSS are forced to decrease. Thus, the criterion will be dominated by the penalty term and the selection of base-learners in later iterations is based solely on the penalty while neglecting the fit to the data. Hence, base-learners with fewer df are preferred, more flexible terms are even completely ignored in later iterations. The same reasoning applies for BIC as well as F-tests. The problem that arises here is that the penalty is not chosen adaptively to the maximum variance that *could* be explained. Thus, unbiased model selection seems not possible if different dfs are specified. As we directly specify the dfs of the base-learners, no high-dimensional optimization via cross validation is needed to choose appropriate smoothing parameters. Choosing the appropriate complexity is reduced to one dimensional cross validation to choose an appropriate stopping iteration m_{stop}. However, due to the iterative nature of boosting, the final degrees of freedom for one covariate can vary greatly even if equal degrees of freedom are specified for the base-learners.

Implementation

The R (R Development Core Team, 2009) add-on package **mboost** (Hothorn *et al.*, 2009) implements component-wise boosting for various loss functions, i.e., for a broad range of regression problems such as Gaussian, binomial or even survival regression models. Structured additive models as discussed in this paper can be fitted using the function `gamboost`. The proposed base-learners are all implemented and can, for example, be applied using the formula interface of `gamboost`. Further base-learners for random effects and spatial effects, among others, are also available.

References

Binder H, Schumacher M (2008). "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models." *BMC Bioinformatics*, **9**, 14.

Breiman L (2001). "Random Forests." *Machine Learning*, **45**(1), 5–32.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and Regression Trees*. Wadsworth, California.

Bühlmann P, Hothorn T (2007). "Boosting Algorithms: Regularization, Prediction and Model Fitting." *Statistical Science*, **22**, 477–505.

Bühlmann P, Yu B (2003). "Boosting with the L_2 Loss: Regression and Classification." *Journal of the American Statistical Association*, **98**, 324–339.

Buja A, Hastie T, Tibshirani R (1989). "Linear Smoothers and Additive Models (with discussion)." *The Annals of Statistics*, **17**, 453–555.

Eilers PHC, Marx BD (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.

Freund Y, Schapire R (1996). "Experiments with a new boosting algorithm." In "Proceedings of the Thirteenth International Conference on Machine Learning," pp. 148–156. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Friedman J, Hastie T, Tibshirani R (2000). "Additive logistic regression: a statistical view of boosting (with discussion)." *The Annals of Statistics*, **28**, 337–407.

Gertheiss J, Tutz G (2009). “Penalized Regression with Ordinal Predictors.” *International Statistical Review*, **77**, 345–365.

Hastie T, Tibshirani R (1986). “Generalized Additive Models.” *Statistical Science*, **1**, 297–310.

Hastie T, Tibshirani R (1990). *Generalized Additive Models*. Chapman & Hall / CRC , London.

Hoerl AE, Kennard RW (1970). “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics*, **12**, 55–67.

Hofner B, Hothorn T, Kneib T (2008). “Variable Selection and Model Choice in Structured Survival Models.” *Technical Report 43*, Department of Statistics, Ludwig-Maximilans-Universität München. URL `http://epub.ub.uni-muenchen.de/7901/`.

Hothorn T, Bühlmann P, Dudoit S, Molinaro A, van der Laan MJ (2006a). “Survival Ensembles.” *Biostatistics*, **7**, 355–373.

Hothorn T, Bühlmann P, Kneib T, Schmid M, Hofner B (2009). *mboost: Model-Based Boosting*. R package version 1.1-4, URL `http://CRAN.R-project.org/package=mboost`.

Hothorn T, Hornik K, Zeileis A (2006b). “Unbiased Recursive Partitioning: A Conditional Inference Framework.” *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.

Kim H, Loh WY (2003). “Classification Trees with Bivariate Linear Discriminant Node Models.” *Journal of Computational and Graphical Statistics*, **12**, 512–530.

Kneib T, Fahrmeir L (2006). “Structured Additive Regression for Categorical Space-Time Data: A Mixed Model Approach.” *Biometrics*, **62**, 109–118.

Kneib T, Fahrmeir L (2010). “A Space-Time Study on Forest Health.” In RE Chandler, M Scott (eds.), “Statistical Methods for Trend Detection and Analysis in the Environmental Sciences,” Wiley. To appear.

Kneib T, Hothorn T, Tutz G (2009). "Variable Selection and Model Choice in Geoadditive Regression Models." *Biometrics*, **65**, 626–634.

Loh WY (2002). "Regression Trees With Unbiased Variable Selection And Interaction Detection." *Statistica Sinica*, **12**, 361–386.

Loh WY, Vanichsetakul N (1988). "Tree-Structured Classification via Generalized Discriminant Analysis." *Journal of the American Statistical Association*, **83**, 715–725. With discussion.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Ruppert D, Wand M, Carroll R (2003). *Semiparametric regression*. Cambridge University Press.

Schmid M, Hothorn T (2008a). "Boosting additive models using component-wise P-splines." *Computational Statistics & Data Analysis*, **53**(2), 298–311.

Schmid M, Hothorn T (2008b). "Flexible Boosting of Accelerated Failure Time Models." *BMC Bioinformatics*, **9**(269).

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007). "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics*, **8**, 25.

Tutz G, Binder H (2006). "Generalized Additive Modelling with Implicit Variable Selection by Likelihood-Based Boosting." *Biometrics*, **62**, 961–971.

van der Laan M (2006). "Statistical inference for variable importance." *The International Journal of Biostatistics*, **2**(1), 1008.

Vapnik V (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, U.S.A.

A Proofs

Proof for Theorem 3.1 The difference of the RSS is given by

$$\begin{aligned}\boldsymbol{\Delta} = \text{RSS}_{\boldsymbol{X}_1} - \text{RSS}_{\boldsymbol{X}_2} &= (\boldsymbol{u} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{u} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1) - (\boldsymbol{u} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2)^\top(\boldsymbol{u} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2)\\ &= \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top)^2\boldsymbol{u} - \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top)^2\boldsymbol{u}\\ &= \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top)\boldsymbol{u} - \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top)\boldsymbol{u}\\ &= \boldsymbol{u}^\top\boldsymbol{Q}\boldsymbol{u}\end{aligned}$$

with $\boldsymbol{Q} = [\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top] - [\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top]$ and

$$\text{tr}(\boldsymbol{Q}) = [n - (M_1 - 1)] - [n - (M_2 - 1)] = M_2 - M_1 \ . \tag{15}$$

Using the theorem for the expected value of quadratic forms (Ruppert *et al.*, 2003, App. A.4.5) it holds:

$$\mathbb{E}(\boldsymbol{\Delta}) = \mathbb{E}(\boldsymbol{u}^\top\boldsymbol{Q}\boldsymbol{u}) = \text{tr}[\boldsymbol{Q}\text{cov}(\boldsymbol{u})] + \mathbb{E}(\boldsymbol{u})^\top\boldsymbol{Q}\mathbb{E}(\boldsymbol{u}) \ . \tag{16}$$

As we assumed $\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0}$ and $\text{cov}(\boldsymbol{u}) = \sigma^2\boldsymbol{I}$, we obtain

$$\mathbb{E}(\boldsymbol{\Delta}) = \sigma^2\,\text{tr}(\boldsymbol{Q}) = \sigma^2(M_2 - M_1) \ . \tag{17}$$

□

Proof for Theorem 3.2 The difference of the RSS in the ridge penalized model is given by

$$\begin{aligned}\boldsymbol{\Delta} &= (\boldsymbol{u} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{\text{pen},1})^\top(\boldsymbol{u} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{\text{pen},1}) - (\boldsymbol{u} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{\text{pen},2})^\top(\boldsymbol{u} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{\text{pen},2})\\ &= (\boldsymbol{u} - \boldsymbol{S}_1\boldsymbol{u})^\top(\boldsymbol{u} - \boldsymbol{S}_1\boldsymbol{u}) - (\boldsymbol{u} - \boldsymbol{S}_2\boldsymbol{u})^\top(\boldsymbol{u} - \boldsymbol{S}_2\boldsymbol{u})\\ &= \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1 + \lambda_1\boldsymbol{K}_1)^{-1}\boldsymbol{X}_1^\top)^2\boldsymbol{u} - \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2 + \lambda_2\boldsymbol{K}_2)^{-1}\boldsymbol{X}_2^\top)^2\boldsymbol{u}\\ &= \boldsymbol{u}^\top\boldsymbol{Q}_{\text{pen}}\boldsymbol{u}\end{aligned}$$

with

$$\begin{aligned}\boldsymbol{Q}_{\text{pen}} &= (\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1 + \lambda_1\boldsymbol{K}_1)^{-1}\boldsymbol{X}_1^\top)^2 - (\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2 + \lambda_2\boldsymbol{K}_2)^{-1}\boldsymbol{X}_2^\top)^2\\ &= -2\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1 + \lambda_1\boldsymbol{K}_1)^{-1}\boldsymbol{X}_1^\top + (\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1 + \lambda_1\boldsymbol{K}_1)^{-1}\boldsymbol{X}_1^\top)^2\\ &\quad +2\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2 + \lambda_2\boldsymbol{K}_2)^{-1}\boldsymbol{X}_2^\top - (\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2 + \lambda_2\boldsymbol{K}_2)^{-1}\boldsymbol{X}_2^\top)^2 \qquad (18)\\ &= \left[-2\boldsymbol{S}_1 + \boldsymbol{S}_1^\top\boldsymbol{S}_1\right] - \left[-2\boldsymbol{S}_2 + \boldsymbol{S}_2^\top\boldsymbol{S}_2\right] \ .\end{aligned}$$

With $\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0}$ and $\text{cov}(\boldsymbol{u}) = \sigma^2 \boldsymbol{I}$, $\sigma^2 > 0$ it follows from (17) that

$$\begin{aligned} \mathbb{E}(\boldsymbol{\Delta}) &= 0 \quad \Leftrightarrow \\ \text{tr}(\boldsymbol{Q}_{\text{pen}}) &= 0 \quad \Leftrightarrow \\ \text{tr}\left\{\left[-2\boldsymbol{S}_1 + \boldsymbol{S}_1^\top \boldsymbol{S}_1\right] - \left[-2\boldsymbol{S}_2 + \boldsymbol{S}_2^\top \boldsymbol{S}_2\right]\right\} &= 0 \quad \Leftrightarrow \\ \text{tr}\left[2\boldsymbol{S}_1 - \boldsymbol{S}_1^\top \boldsymbol{S}_1\right] &= \text{tr}\left[2\boldsymbol{S}_2 - \boldsymbol{S}_2^\top \boldsymbol{S}_2\right] \ . \end{aligned}$$

□

B Web Supplement - Theorem with Proof for Smooth Base-Learners

Theorem B.1 *Let $\boldsymbol{x}$ be a continuous covariate with design matrix $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x})$. A smooth effect for $\boldsymbol{x}$ is modeled using P-splines with the design matrix $\boldsymbol{B} = (B_1(\boldsymbol{x}), \ldots, B_k(\boldsymbol{x}))$, which consists of B-spline basis functions evaluated at the values of $\boldsymbol{x}$. Let $\boldsymbol{K} = \boldsymbol{D}^\top \boldsymbol{D}$ be the penalty matrix, where $\boldsymbol{D}$ is a difference matrix of order 2. The corresponding smoothing parameter is denoted by λ. Let $\boldsymbol{u}$ be the $n \times 1$ negative gradient vector arising in the first step of the boosting algorithm for a response variable $\boldsymbol{y}$ of i.i.d. normally distributed random variables with variance σ^2 that is independent of $\boldsymbol{x}$, i.e. $\boldsymbol{u}$ is simply the centered response variable. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{\text{pen}}$ denote the effect estimates resulting from unpenalized and penalized least squares base-learners and define the difference of the residual sum of squares as $\boldsymbol{\Delta} = (\boldsymbol{u} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{u} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - (\boldsymbol{u} - \boldsymbol{B}\hat{\boldsymbol{\beta}}_{\text{pen}})^\top(\boldsymbol{u} - \boldsymbol{B}\hat{\boldsymbol{\beta}}_{\text{pen}})$. Then it holds that*

$$\mathbb{E}(\Delta) > 0 \quad (\textit{if } \lambda < \infty). \tag{19}$$

Proof for Theorem B.1 The difference of the RSS is given by

$$\begin{aligned} \boldsymbol{\Delta} = \text{RSS}_{\text{lin}} - \text{RSS}_{\text{pen}} &= \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^2\boldsymbol{u} - \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{K})^{-1}\boldsymbol{B}^\top)^2\boldsymbol{u} \\ &= \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{u} - \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{K})^{-1}\boldsymbol{B}^\top)^2\boldsymbol{u} \\ &= \boldsymbol{u}^\top\boldsymbol{Q}\boldsymbol{u} \end{aligned}$$

with

$$
\begin{aligned}
\boldsymbol{Q} &= [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top] - [\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{K})^{-1}\boldsymbol{B}^\top]^2 \\
&= \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + 2\boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{K})^{-1}\boldsymbol{B}^\top - (\boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{K})^{-1}\boldsymbol{B}^\top)^2. \qquad (20)
\end{aligned}
$$

It holds that $\mathrm{tr}(\boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{K})^{-1}\boldsymbol{B}^\top) = \mathrm{tr}((\boldsymbol{I} + \lambda \widetilde{\boldsymbol{K}})^{-1}) = \sum_{j=1}^k (1 + \lambda d_j)^{-1}$, where $\widetilde{\boldsymbol{K}} = (\boldsymbol{B}^\top \boldsymbol{B})^{-1/2}\boldsymbol{K}(\boldsymbol{B}^\top \boldsymbol{B})^{-1/2}$ and d_j are the eigenvalues of $\widetilde{\boldsymbol{K}}$ (cf. Eilers and Marx, 1996). For second order difference penalty matrices $\boldsymbol{D}$ two eigenvalues are equal to zero, all others eigenvalues d_j are positive. Thus

$$
\begin{aligned}
\mathrm{tr}(\boldsymbol{Q}) &= -2 + 2\sum_{j=1}^k (1 + \lambda d_j)^{-1} - \sum_{j=1}^k (1 + \lambda d_j)^{-2} \\
&\geq -2 + 2\sum_{j=1}^k (1 + \lambda d_j)^{-1} - \sum_{j=1}^k (1 + \lambda d_j)^{-1} \\
&= -2 + \sum_{j=1}^k (1 + \lambda d_j)^{-1} \geq 0,
\end{aligned}
$$

where $\mathrm{tr}(\boldsymbol{Q}) = 0$ if and only if $\lambda \to \infty$. As we assumed $\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0}$ and $\mathrm{cov}(\boldsymbol{u}) = \sigma^2 \boldsymbol{I}$, by using (16) we obtain for a finite smoothing parameter λ

$$
\mathbb{E}(\boldsymbol{\Delta}) = \sigma^2\, \mathrm{tr}(\boldsymbol{Q}) > 0.
$$

□

From (20) we can see that $\mathrm{tr}(2\boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{K})^{-1}\boldsymbol{B}^\top - (\boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{K})^{-1}\boldsymbol{B}^\top)^2) = \mathrm{tr}(2\boldsymbol{S} - \boldsymbol{S}^\top \boldsymbol{S})$ needs to be controlled to make the terms comparable. Thus, the appropriate degrees of freedom have the same form as for categorical effects.

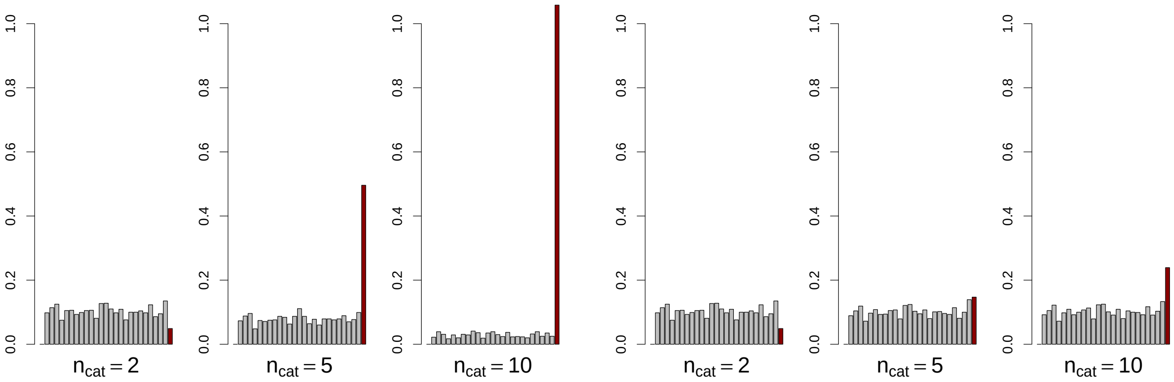
C Web Supplement - Model for Forest Health Data

In the application we modeled the health status of beeches in a northern Bavarian forest district. More details on the data and the results can be found in Section 5 of the paper. An overview of the covariates can be found in Table 2. Based on previous analyses (e.g., Kneib and Fahrmeir, 2006; Kneib *et al.*, 2009; Kneib and Fahrmeir, 2010) we consider a candidate model with the additive predictor

$$\begin{aligned} \eta &= \boldsymbol{x}^\top \boldsymbol{\beta} + f_1(\text{ph}) + f_2(\text{canopy}) + f_3(\text{soil}) + f_4(\text{inclination}) \\ &\quad + f_5(\text{elevation}) + f_6(\text{time}) + f_7(\text{age}) + f_8(s_1, s_2) + b_{\text{plot}}, \end{aligned} \tag{21}$$

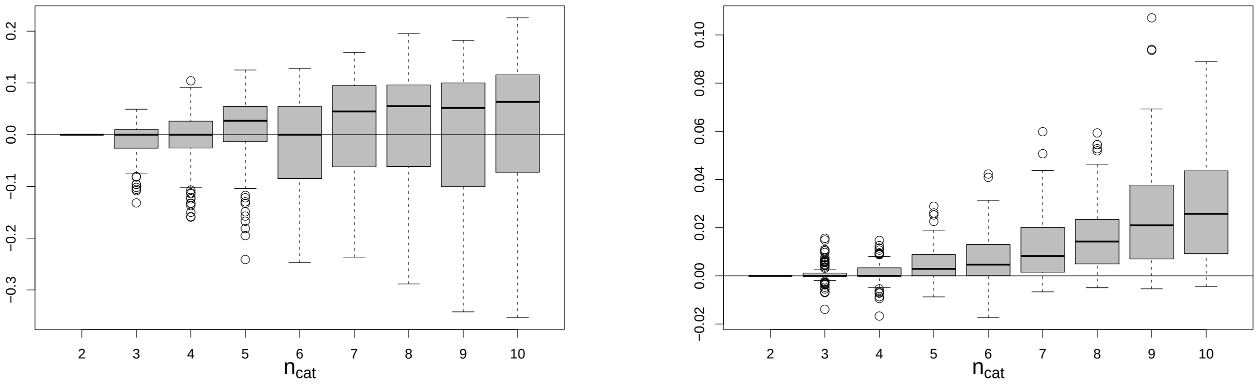
where $\boldsymbol{x}$ contains the parametric effects of the categorical covariates fertilization, stand, humus and saturation. The ordinal covariates humus and saturation are modeled using ridge penalized ordinal base-learners, whereas the other categorical covariates are modeled using ridge penalized base-learners. The smooth effects $f_1, \dots, f_7$ are specified as a combination of linear base-learners and univariate cubic penalized splines with 20 inner knots and second order difference penalty. For the spatial effect f_8 we assumed bivariate cubic penalized splines with first order difference penalties and 12 inner knots for each of the directions. Finally, the plot-specific effects b_{plot} is represented by a ridge-type "random effects" base-learner with fixed degrees of freedom (see Kneib *et al.*, 2009, for details). All continuous covariates were centered. To correct for the selection bias, one degree of freedom is assigned to each single base-learner including the spatial and random effect base-learners.

For the benchmark of the bias corrected model to an uncorrected model we used the same model but specified four degrees of freedom for all smooth base-learners and added the categorical covariates unpenalized.



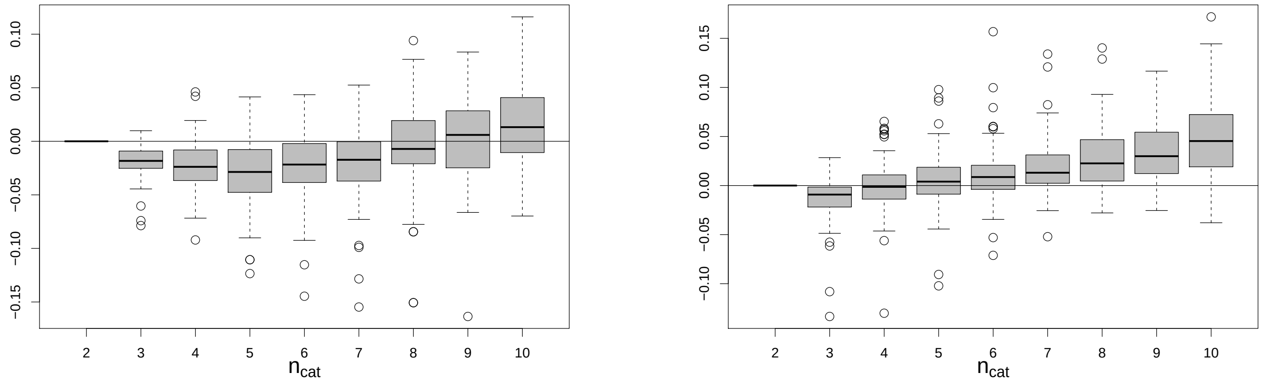
(a) Unpenalized Base-Learner (b) Ridge Penalized Base-Learner

Figure 1: **Null Model:** Average selection frequencies of base-learners for $n_{\text{cat}} = \{2, 5, 10\}$ in the "optimal step" $\widehat{m}_{\text{stop,opt}}$ without and with ridge penalty. The last bar in each graph represents the selection frequency of the categorical covariate.



(a) Difference of Relative Selection Frequencies (b) $\text{MSE}_{\text{unpenalized}} - \text{MSE}_{\text{penalized}}$

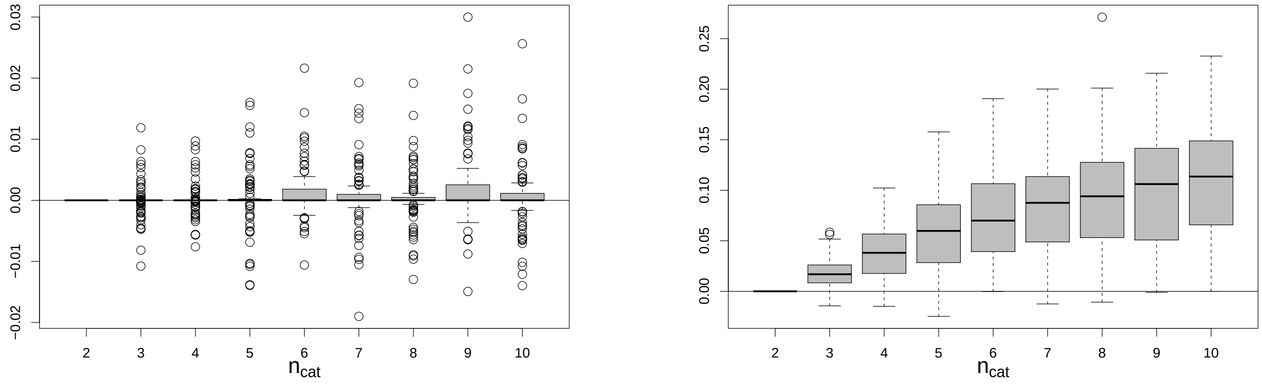
Figure 2: **Power Case 1:** Differences in relative number of selections for categorical base-learner (unpenalized - penalized) (left) and differences of MSE (right).



(a) MSE with z_2

(b) MSE without z_2

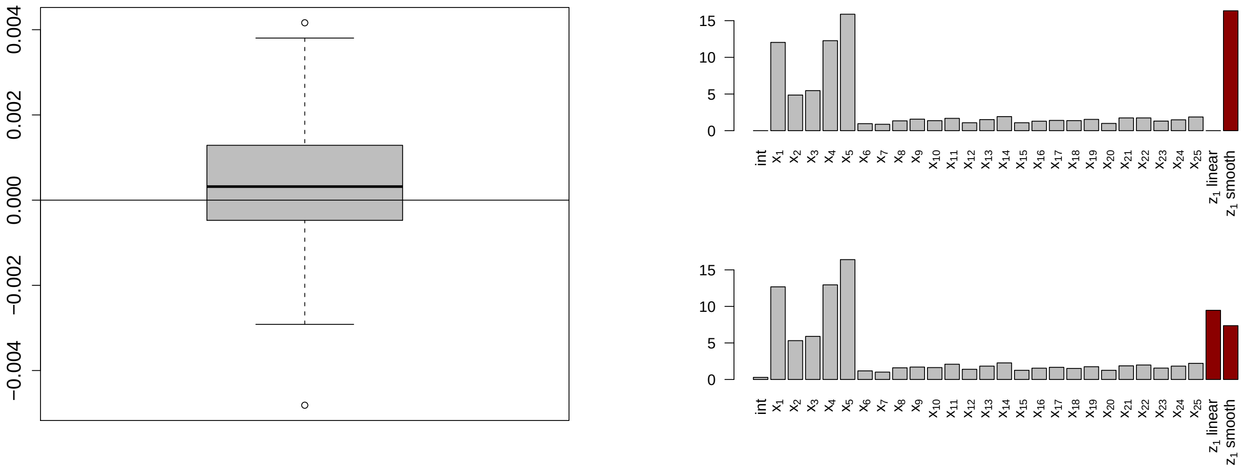
Figure 3: **Power Case 2:** Boxplots represent $\text{MSE}_{\text{unpenalized}} - \text{MSE}_{\text{penalized}}$ where the MSE is computed with (left) and without (right) the influential, categorical covariate z_2.



(a) Power Case 1

(b) Power Case 2

Figure 4: **Comparison of model with ridge penalized and ordinal penalized base-learner:** Boxplots represent $\text{MSE}_{\text{ridge penalized}} - \text{MSE}_{\text{ordinal penalized}}$ with non-influential categorical covariate (left) and with additional influential categorical covariate (right).



(a) Partial Deviation

(b) Selection Frequency

Figure 5: **Power Case 2 (Continuous Covariate):** Partial deviation $\Delta^{L_2}_{\text{model}} - \Delta^{L_2}_{\text{model with decomposition}}$ (left) and mean selection frequency of base-learners in the "optimal step" $\widehat{m}_{\text{stop,opt}}$ with 4 df (upper) and decomposition (i.e., 1 df; lower). The last bar in each graph represents the smooth term and the second bar from the right represents the linear term (for z_1). The bars 2 to 6 represent the influential covariates $x_1, \ldots, x_5$. Note that the height of the bar increases with increasing effect size $|\beta_i|$.

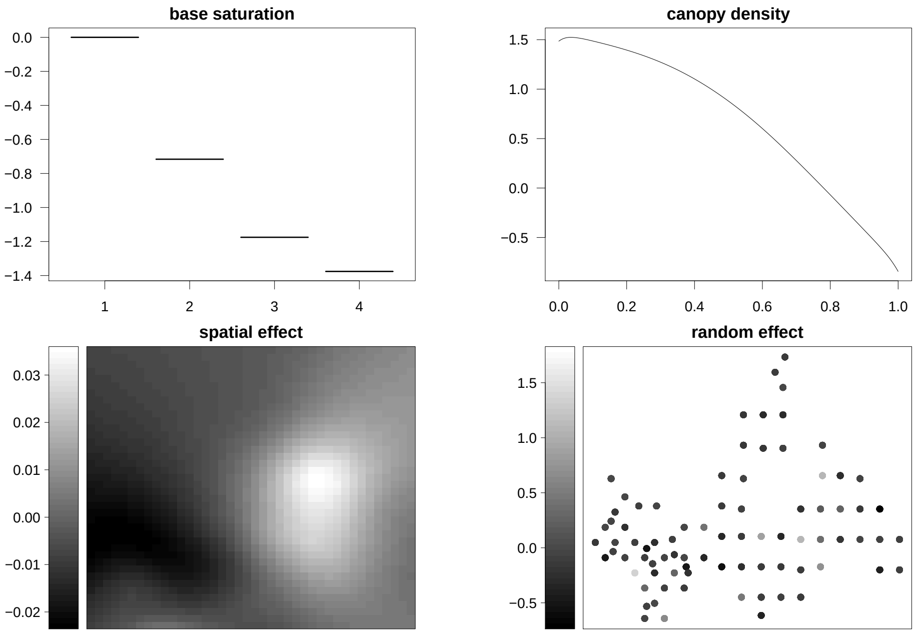


Figure 6: **Forest Health:** Effects of base saturation, canopy density, spatial effect and random effect (without spatial variation).

Table 1: **Simulation:** Overview of different schemes for categorical covariates z_1 and z_2 (upper part) and continuous covariate z_1 (lower part)

		Effects for $x_1, \ldots, x_{25}$	Effects for z_1 (and z_2)
Null Model		$\boldsymbol{\beta} = (0, \ldots, 0)^\top$	$\boldsymbol{\gamma}_1 = (0, \ldots, 0)^\top$
Power Case 1	z_1 non-influential	$\boldsymbol{\beta} = (-2, -1, 1, 2, 3, 0, \ldots, 0)^\top$	$\boldsymbol{\gamma}_1 = (0, \ldots, 0)^\top$
Power Case 2	z_1 non-influential	$\boldsymbol{\beta} = (-2, -1, 1, 2, 3, 0, \ldots, 0)^\top$	$\boldsymbol{\gamma}_1 = (0, \ldots, 0)^\top$
	z_2 influential		$\boldsymbol{\gamma}_2 = (\frac{2}{n_{\text{cat}}/2}, \frac{3}{n_{\text{cat}}/2}, \ldots, \frac{n_{\text{cat}}}{n_{\text{cat}}/2})^\top$
Null Model		$\boldsymbol{\beta} = (0, \ldots, 0)^\top$	$f_z(z_1) \equiv 0$
Power Case 1	z_1 non-influential	$\boldsymbol{\beta} = (-2, -1, 1, 2, 3, 0, \ldots, 0)^\top$	$f_z(z_1) \equiv 0$
Power Case 2	linear effect of z_1	$\boldsymbol{\beta} = (-2, -1, 1, 2, 3, 0, \ldots, 0)^\top$	$f_z(z_1) = 1.5z_1$
Power Case 3	smooth effect of z_1	$\boldsymbol{\beta} = (-2, -1, 1, 2, 3, 0, \ldots, 0)^\top$	$f_z(z_1) = \sin(-(2z_1)^2 - 0.6(2z_1)^3)$

Table 2: **Forest health data:** Description of covariates. All continuous covariates were centered before included in the model, categorical covariates are dummy coded with the first category as reference.

Covariate	Description
age	average age of trees at the observation plot in years (continuous, $7 \leq \text{age} \leq 234$)
time	calendar time (continuous, $1983 \leq \text{time} \leq 2004$)
elevation	elevation above sea level in meters (continuous, $250 \leq \text{elevation} \leq 480$)
inclination	inclination of slope in percent (continuous, $0 \leq \text{inclination} \leq 46$)
soil	depth of soil layer in centimeters (continuous, $9 \leq \text{soil} \leq 51$)
ph	ph-value at 0-2cm depth (continuous, $3.28 \leq \text{ph} \leq 6.05$)
canopy	density of forest canopy in percent (continuous, $0 \leq \text{canopy} \leq 1$)
humus	thickness of humus layer in 5 categories (ordinal, higher categories represent higher proportions)
saturation	base saturation in 4 categories (ordinal, higher categories indicate higher base saturation)
stand	type of stand (categorical, -1 = mixed forest, 1 = deciduous forest)
fertilization	fertilization (categorical, -1 = no, 1 = yes)