Selective Inference for $L_2$-Boosting

David Rügamer
and
Sonja Greven
Department of Statistics, LMU Munich

October 30, 2018

Abstract

We adapt and extend several recent proposals for post-selection inference to transfer them to the component-wise functional gradient descent algorithm (CFGD) under normality assumption for model errors, also known as $L_2$-Boosting. The CFGD is one of the most versatile toolboxes to analyze data as it scales well to high-dimensional data sets, allows for a very flexible definition of additive regression models and incorporates inbuilt variable selection. Due to the iterative nature, which can repeatedly select the same component to update, a statistical inference framework for component-wise boosting algorithms requires adaptations of existing approaches; we propose tests and confidence intervals for linear, grouped and penalized additive model components selected by $L_2$-Boosting. Our concepts also transfer to slow-learning algorithms and to other selection techniques which restrict the response space to more complex sets than polyhedra. We apply our framework to an additive model for the prostate cancer data set to compare with previous results, and investigate the properties of our concepts in simulation studies.

Keywords: Bootstrap, Functional Gradient Descent Boosting, Post-Selection Inference, Selective Inference, Slow Learner
1 Introduction

**Inference for boosting.** In this work we review and adapt recently proposed statistical inference techniques to the component-wise functional gradient descent algorithm (CFGD; see, e.g., Hothorn et al. 2010). CFGD emerged from the field of machine learning (c.f. Friedman 2001), but has since also become an algorithm used to estimate statistical models (see, e.g., Mayr et al. 2017, Melcher et al. 2017, Rügamer et al. 2018, Brockhaus et al. 2018). A commonly used and well studied special CFGD algorithm is $L_2$-Boosting (Bühlmann & Yu 2003). Apart from Luo & Spindler (2017), who study uncertainty for treatment effects when selecting control variables via $L_2$-Boosting in instrumental variable models, which requires additional assumptions for all the variables in the model, no general inferential concepts in the sense of classical statistical inference have been proposed for $L_2$-Boosting yet, though ad-hoc solutions such as a non-parametric bootstrap are often used to quantify the uncertainty of boosting estimates (see e.g. Brockhaus et al. 2015, Rügamer et al. 2018).

In many research areas such an uncertainty quantification is indispensable. We therefore propose a framework for conducting valid inference for regression coefficients in models fitted with $L_2$-Boosting by conditioning on the selected covariates. We adapt recent research findings on selective inference, which transfers classical statistical inference to algorithms that rely on a preceding selection of model terms as is the case for CFGD algorithms. Compared to existing approaches for sequential regression procedures including forward stepwise regression (Tibshirani et al. 2016), inference for $L_2$-Boosting carries additional challenges due to an iterative procedure that can repeatedly select the same model term. In contrast to methods, for which selective inference is already available, our method additionally allows for the estimation of non-linear effects. We demonstrate this flexibility by applying the proposed framework to the prostate cancer dataset, which has been used to demonstrate selective inference after the Lasso selection, however, with the restriction of linearity of all covariate effects.

**Suitable inference concepts.** The invalidity of inference after model selection has been mentioned by many authors throughout the last decades (see, e.g., Berk et al. 2013) and leads to the necessity for a suitable inference framework. Different approaches for inference
in high-dimensional regression models have emerged over the past years, including data splitting \cite{WassermanRoeder2009}. Apart from these techniques, post-selection inference (PoSI; \cite{Berketal2013}) attracts growing interest. Initiated by the proposal for valid statistical inference after arbitrary selection procedures by \cite{Berketal2013}, many new findings and adoptions of post-selection inference to known statistical methods have been published in the last years.

We here focus on selective inference, which provides inference statements conditional on the observed model selection. Similar to data splitting, selective inference separates the information in the data, which is used for the model selection, from the information, which is used to infer about parameters post model selection. In contrast to the original PoSI idea of providing simultaneous inference for every possible model selection, selective inference is designed to yield less conservative inference statements.

On the downside, selective inference can yield unstable and potentially infinite confidence intervals in certain situations. This was recently shown by \cite{KivaranovicLeeb2018} for selective inference concepts based on polyhedral constraints. However, for our method exploiting that the selective space is a union of polyhedra, this seems to be rarely the case. Our simulation studies show powerful inference despite settings with low signal-to-noise ratio and/or with the number of predictors exceeding the number of observations prior to model selection. This suggests that using the same approach for the Lasso selection when not conditioning on a list of signs, which also results in a union of polyhedra, might help alleviate this problem and lead to more powerful inference.

Apart from general theory described in \cite{Fithianetal2014}, which transfers the classical theory to selective inference in exponential family models following any type of selection mechanism, different explicit selective inference frameworks for several selection methods have been derived (see e.g. \cite{Leeetal2016}, for selective inference after Lasso selection or \cite{RugamerGreven2018}, for selective inference after likelihood- and test-based model selection). Recent publications, which are particularly relevant for this work, aim for valid inference in forward stepwise regression \cite{Tibshiranietal2016,LoftusTaylor2014,LoftusTaylor2015}. Whereas \cite{Tibshiranietal2016} build a framework for any sequential regression technique resulting in a limitation to the space for inference, which can be characterized by a poly-
hedral set, Loftus & Taylor (2014, 2015) extend the idea to a more general framework, for which the inference space is given by quadratic inequalities and which coincides with the polyhedral approach in special cases. A continuation of Loftus & Taylor (2015) is given by Yang et al. (2016). With the objective to build a selective inference framework for the group Lasso (Yuan & Lin 2006), Yang et al. describe an importance sampling algorithm that circumvents the problem of having to explicitly define the space, to which the inference is restricted after conditioning on the selected model.

**Resampling for uncertainty quantification.** Uncertainty quantification by the use of resampling methods is as error-prone as classical inference when applied to models after a certain model selection procedure. We therefore shortly address this issue by the example of bootstrap as one of the most commonly used techniques.

Let us first consider the parametric bootstrap. When generating new samples of the response from the selected model and conducting unadjusted inference, the selected model is treated as the true model and this can incorrectly lead effects to be (non-)zero. A non-parametric bootstrap on the other hand is accompanied by its own problems. First, when drawing pairs of response and covariates, we (implicitly) assume that the underlying data model is based on a random design in contrast to many regression model settings, where the covariates are assumed to be fixed. If we ignore this issue, we still face the problem of either neglecting the uncertainty of model selection, if we refit the initially selected model for the resampled data, or the problem of having to aggregate over different models when integrating the model selection process into our resampling procedure. If estimates are aggregated over different models, uncertainty quantification of parameters is based on different selected models with different interpretations of the estimated coefficients based on projections of the mean into different subspaces. This quantifies variability of estimates over the selected but not over all possible models. In particular, small true effects might never be selected, yielding a zero confidence interval with no proper coverage. An additional difficulty arises when using the bootstrap for boosted regression models, in which the estimated coefficients exhibit a bias due to the shrinkage effect of boosting. Hence, bootstrap intervals are not centered around the true value and thus yield a quantification
of variability rather than a measure of deviation from the truth.

**Contribution of this work.** In this work, we adapt and extend several existing approaches for selective inference, thereby addressing the following issues: 1. We explicitly derive the space restriction of the response given by the $L_2$-Boosting path and thereby allow for inference as proposed in [Tibshirani et al. (2016)](Tibshirani2016). 2. We propose a new conditional inference concept for $L_2$-Boosting and potentially other slow learning algorithms by conditioning on a set of possible selection paths. This idea can also be used to conduct inference for inference problems, where one has to condition on additional quantities in order to facilitate explicit calculations. This, for example, is the case when inference is sought for the Lasso and an analytic representation of the inference space only becomes feasible after additionally conditioning on a list of signs. 3. Computation of p-values and (two-sided) confidence intervals is done by Monte Carlo approximation following results of [Tibshirani et al. (2016)](Tibshirani2016) and [Yang et al. (2016)](Yang2016). This circumvents an explicit mathematical representation of the space the test statistic is truncated to. While we apply this approach to $L_2$-Boosting, for which the test statistic lies in a union of polyhedra and therefore has a (conditional) normal distribution with potentially multiple truncation limits, our framework does not assume a certain type of space restriction induced by the model selection procedure. It is thus applicable whenever the model of interest is of additive nature and the response variable is assumed to be normally distributed. 4. We explain how the proposed inference concept can easily be extended to account for cross-validation, stability selection ([Shah & Samworth 2013](Shah2013)) and similar sub-sampling methods. 5. We extend the idea of the selective inference framework to models including $L_2$-penalized additive effects, such as smooth effects.

In the following, we describe the $L_2$-Boosting algorithm in section 2 and recapitulate the concept of selective inference for sequential regression procedures in section 3. In section 4 we investigate the challenges accompanying a new inference framework for $L_2$-Boosting and propose several solutions. In section 5 we present simulation results and analyze the prostate cancer data using our new approach in section 6. We discuss limitations and further extensions of the approach in section 7. An add-on R-package to the model-based
boosting R package mboost is available at https://github.com/davidruegamer/iboost, which can be used to conduct inference for boosted models and to reproduce the results of section 5 and 6. Further simulation and applications results as well as code to reproduce the simulation results are given in an online appendix.

2 \(L_2\)-Boosting

We now present the \(L_2\)-Boosting algorithm as a special generic CFGD algorithm. Let \(X \in \mathbb{R}^{n \times p}\) be a fixed set of covariates and \(y\) a realization of the random response variable \(Y \in \mathbb{R}^n\). The goal is to minimize a loss function \(\ell(\cdot, y)\) for the given realization \(y\) with respect to an additive model \(f := \sum_{j=1}^J g_j(X_j)\), where function evaluations of \(g_j\) are evaluated row-wise. The functions \(g_j(\cdot)\), the so called base-learners, are defined for column subsets \(X_j \in \mathbb{R}^{n \times p_j}\) of \(X\) with \(1 \leq p_j \leq p\) and can be fitted to some vector \(u \in \mathbb{R}^n\), which yields \(\hat{g}_j\) as estimate for \(g_j(X_j)\). We estimate \(f\) by \(\hat{f}\) using the component-wise functional gradient descent algorithm:

1. Initialize an offset value \(\hat{f}^{(0)} \in \mathbb{R}^n\). If \(y\) is centered, a natural choice is \(\hat{f}^{(0)} = (0, \ldots, 0)^\top\). Define \(m = 0\).
2. Do the following for \(m = 1, \ldots, m_{\text{stop}}\):
   1. Compute the pseudo-residuals \(u^{(m)} \in \mathbb{R}^n\) of step \(m\) as \(u^{(m)} = -\frac{\partial}{\partial f} \ell(f, y)\bigg|_{f=\hat{f}^{(m-1)}}\).
   2. Approximate the negative gradient vector \(u^{(m)}\) with \(\hat{g}_j\) by fitting each of the base-learners \(g_j(\cdot), j = 1, \ldots, J\) to the pseudo-residuals and find the base-learner \(j^{(m)}\), for which \(j^{(m)} = \arg\min_{1 \leq j \leq J} \|u^{(m)} - \hat{g}_j\|_2^2\) holds.
   3. Update \(\hat{f}^{(m)} = \hat{f}^{(m-1)} + \nu \cdot \hat{g}_{j^{(m)}}\), where \(\nu \in (0, 1]\) is the so called step-length or learning rate and usually fixed to some sufficiently small value such as 0.1 or 0.01 (Bühlmann & Hothorn 2007).

When defining \(\ell(f, y) = \frac{1}{2}\|y - f\|_2^2\) with quadratic \(L_2\)-Norm \(\|\cdot\|_2^2\), \(L_2\)-Boosting is obtained, which corresponds to mean regression using the model \(E(Y | X) = \sum_{j=1}^J g_j(X_j)\). The vector \(u^{(m)}\) then corresponds to the residuals \(y - \hat{f}^{(m-1)}\). In the framework of additive regression models, each base-learner \(g_j(\cdot)\) constitutes a partial effect and is represented as linear effect.
of a covariate or of a basis evaluated at that covariate vector, i.e., \( g_j(X_j) = X_j \beta_j \). \( \beta_j \) is estimated using ordinary or penalized least squares. The model fit \( \hat{g}_j^{(m)} \) of each base-learner in the \( m \)th step is therefore given by \( \hat{g}_j^{(m)} = H_j u^{(m)} = X_j (X_j^\top X_j + \lambda_j D_j)^{-1} X_j^\top u^{(m)} \), where the hat matrix \( H_j \) is defined by the corresponding design matrix \( X_j \), a penalty matrix \( D_j \) and a pre-specified smoothing parameter \( \lambda_j \geq 0 \) controlling the penalization. As only one base-learner is chosen in each iteration, the final effective degrees of freedom of the \( j \)th base-learner depend on the number of selections.

As \( L_2 \)-Boosting scales well to large data sets due to its component-wise fitting nature and is particularly suited for the estimation of structured additive regression models, it is often used as an estimation algorithm for a statistical additive model (see, e.g., [Mayr et al., 2017]). It has the additional advantage of being able to handle \( n < p \)-settings and conducting variable selection, as not all \( J \) model terms are necessarily selected in at least one iteration. However, when constructing a measure of uncertainty for regression coefficients, the preceding variable selection has to be accounted for. As for other variable selection procedures, the iterative nature of \( L_2 \)-Boosting restricts the space of \( Y \) and thereby the space of estimated parameters.

### 3 Selective Inference for Sequential Regression Procedures

We first define the considered model framework and some necessary notations (Section 3.1) before reviewing existing selective inference approaches (Section 3.2 – 3.4) we build on in Section 4.

#### 3.1 Considered Setup

Let \( Y = \mu + \epsilon \) with \( \epsilon \sim \mathcal{N}(0, \sigma^2 I_n) \) and \( n \)-dimensional identity matrix \( I_n \). Furthermore, assume that \( \sigma^2 \) is known and \( \mu \) is an unknown parameter of interest. In particular, we do not assume any true linear relationship between \( \mu \) and covariates, but estimate \( \mu \) with a “working model”, which is of additive nature based on fixed covariates \( X \in \mathbb{R}^{n \times p} \), for which \( p \) potentially exceeds \( n \). Furthermore, define the selection procedure or selection event \( S \):
\[ \mathbb{R}^n \to \mathcal{P}(\{1, \ldots, p\}), \mathbf{y} \mapsto \mathcal{S}(\mathbf{y}) \text{ with power set function } \mathcal{P}(\cdot). \] For the given realization \( \mathbf{y} \) of \( \mathbf{Y} \), we denote \( \mathcal{S}(\mathbf{y}) := \mathcal{A} \), for which we assume \( |\mathcal{A}| \leq n \).

We focus on estimating the best linear projection of \( \mathbf{\mu} \) into the space spanned by the variables given by \( \mathcal{A} \) after model selection. We therefore run the selection procedure defined by \( \mathcal{S} \), select the subset \( \mathbf{X}_\mathcal{A} \) of \( \mathbf{X} \) defined by the selected column indices \( \mathcal{S}(\mathbf{y}) = \mathcal{A} \) and estimate regression coefficients \( \mathbf{\beta}_\mathcal{A} \) by projecting \( \mathbf{y} \) into the linear subspace \( \mathbf{W}_\mathcal{A} \subseteq \mathbb{R}^n \) spanned by the columns of \( \mathbf{X}_\mathcal{A} \). With the goal to infer about \( \beta_j, j \in \mathcal{A} \), in \( \mathbf{\beta}_\mathcal{A} \), we test the hypothesis \( H_0 : \beta_j = 0 \). This is equivalent to testing

\[
H_0 : \mathbf{v}^\top \mathbf{\mu} := \mathbf{e}_j^T (\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1} \mathbf{X}_\mathcal{A}^\top \mathbf{\mu} = 0
\]

with \( \mathbf{e}_j \) the unit vector selecting \( j \in \mathcal{A} \) (see, e.g., [Tibshirani et al. 2016]).

In a classical statistical approach without selection, \( \hat{\mathbf{R}} := \mathbf{v}^\top \mathbf{Y} \), which follows a normal distribution with expectation \( \hat{\rho} = \mathbf{v}^\top \mathbf{\mu} \) and variance \( \sigma^2 \mathbf{v}^\top \mathbf{v} \) under the null. However, after model selection, the space of \( \mathbf{Y} \) is restricted to \( \mathcal{G} = \{ \mathbf{y} : \mathcal{S}(\mathbf{y}) = \mathcal{A} \} \), which we call the inference region.

Many of the proposed methods for selective inference then describe this space restriction mathematically and derive the distribution of \( \mathbf{v}^\top \mathbf{Y} \mid \mathbf{Y} \in \mathcal{G} \). Let \( \mathbf{P}_\mathbf{W} \) be the projection onto a linear subspace \( \text{span}(\mathbf{W}) \subseteq \mathbb{R}^n \) defined by \( \mathbf{W} \in \mathbb{R}^{n \times w}, w \in \mathbb{N} \) and \( \mathbf{P}_{\mathbf{W}^\perp} \) be the projection onto the orthogonal complement of this linear subspace. Furthermore, define the direction of \( \mathbf{P}_{\mathbf{W}} \mathbf{y} \) as the unit vector \( \text{dir}_\mathbf{W}(\mathbf{y}) = \frac{\mathbf{P}_{\mathbf{W}} \mathbf{y}}{||\mathbf{P}_{\mathbf{W}} \mathbf{y}||^2} \).

We now shortly review three approaches to selective inference derived for a similar setup and build on these ideas in Section 4.

### 3.2 Inference based on a Polyhedral Space Characterization

For sequential regression procedures such as Forward Stepwise Regression (FSR) or the Least Angle Regression (LAR, [Efron et al. 2004], [Tibshirani et al. 2016]) characterize the restricted region of the on-going selection mechanism as a polyhedral set \( \mathcal{G} = \{ \mathbf{y} : \Gamma \mathbf{y} \geq \mathbf{b} \} \) with \( \Gamma \in \mathbb{R}^{\kappa \times n}, \mathbf{b} \in \mathbb{R}^\kappa \) for some \( \kappa \in \mathbb{N} \) and an inequality \( \geq \) which is to be interpreted componentwise. In their framework \( \mathbf{Y} \) is written as \( \hat{\mathbf{R}} \cdot \frac{\mathbf{v}}{\mathbf{v}^\top \mathbf{v}} + \mathbf{Z} \) with \( \mathbf{Z} = \mathbf{P}_{\mathbf{v}^\perp} \mathbf{Y} \). By
construction $Z$ is independent of $\tilde{R}$. The selection event $Y \in G$ can thus be rewritten
\[ G = \{ Y \text{ with } V_{\text{lo}}(Z) \leq \tilde{R} \leq V_{\text{up}}(Z), V^0(Z) \geq 0 \}, \] (2)
where $V_{\text{lo}}$, $V_{\text{up}}$ and $V^0$ are functions of $Z$ as well as of the fixed quantities $\Gamma$ and $v$. By additionally conditioning on the realization $z$ of $Z$ as well as on a list of signs for each step similar to those defined in (10) and which will be explained in Section 4, $V_{\text{lo}}, V_{\text{up}}$ are fixed limits for $\tilde{R}$ (see, e.g., Lee et al. 2016) with $Y \in G$ corresponding to $\tilde{R} \in \mathbb{R}_{y} := \{ \tilde{R} : V_{\text{lo}}(z) \leq \tilde{R} \leq V_{\text{up}}(z) \}$. Incorporating these boundaries into the distribution of $\tilde{R} \sim \mathcal{N}(\tilde{\rho}, \sigma^2 v^\top v)$ yields a truncated Gaussian distribution with truncation limits $V_{\text{lo}} = V_{\text{lo}}(z), V_{\text{up}} = V_{\text{up}}(z)$. Let $F_{\tilde{R}, \rho, \sigma^2 v^\top v}(\tilde{R})$ denote the cumulative distribution function of this truncated normal distribution evaluated at $\tilde{R}$. Then, for $H_0 : \tilde{\rho} \leq 0$ vs. $H_1 : \tilde{\rho} > 0$, the test statistic $T = 1 - F_{\rho, \sigma^2 v^\top v}(\tilde{R})$ is a valid conditional p-value, conditional on the polyhedral selection, as $\mathbb{P}_{H_0}(T \leq \alpha \mid \Gamma Y \geq b) = \alpha$ for any $0 \leq \alpha \leq 1$. For a two-sided hypothesis $H_0 : \tilde{\rho} = 0$ vs. $H_1 : \tilde{\rho} \neq 0$, Tibshirani et al. (2016) define $T = 2 \cdot \min \left( F_{\rho, \sigma^2 v^\top v}(\tilde{R}), 1 - F_{\rho, \sigma^2 v^\top v}(\tilde{R}) \right)$ and the validity of inference based on this p-value holds analogously.

### 3.3 Model Selection Procedures as Affine Inequalities

The characterization of the inference region as a polyhedral set, however, is only possible if the algorithmic decision in each selection step is a linear restriction on the space of $Y$. For example for groups of variables, the underlying inequality for the choice of the covariate is inherent quadratic and no polyhedral representation can be obtained. Loftus & Taylor (2015) therefore introduce a framework for inference after model selection procedures which can be described by affine inequalities.

Apart from a different characterization of the space restriction, a different test statistic must be used for groups of variables. For testing the $j$th group variable coefficient $\beta_{A,j} \in \mathbb{R}^w$ in the best linear approximation $\beta_A = \arg \min \mathbb{E} [||Y - X_A \beta||^2_2]$, Loftus & Taylor (2015), Yang et al. (2016) rewrite the null hypothesis $\beta_{A,j} = 0$ as $P_W \mu = 0 \iff \rho := ||P_W \mu||_2 = 0$.

\[ H_0 : \rho := ||P_W \mu||_2 = 0 \] (3)
with \( \mathbf{W} = \mathbf{P}^\perp_{\mathbf{X}_{\mathcal{A}\setminus j}} \mathbf{X}_j \), where \( \mathbf{X}_{\mathcal{A}\setminus j} \) denotes \( \mathbf{X}_\mathcal{A} \) without the \( p_j \) columns corresponding to the \( j \)th group variable. In other words we want to test the correlation of \( \mathbf{X}_j \) and \( \mathbf{\mu} \) after adjusting for all other predictors \( \mathcal{A}\setminus j \) in the selected model \( \mathcal{A} \). Using \( R := ||\mathbf{P}_\mathbf{W}\mathbf{Y}||_2 \) as test statistic, the authors then conduct inference. Under the null and when additionally conditioning on the direction \( \text{dir}_{\mathbf{W}}(\mathbf{y}) \), \( R \) follows a truncated \( \chi^2 \)-distribution and truncation limits of \( R \) can again be derived analytically. With the goal to also facilitate the computation of confidence intervals, Yang et al. (2016) note that \( R \) and \( \text{dir}_{\mathbf{W}}(\mathbf{y}) \) are not independent for \( \rho \neq 0 \) and as a consequence, the \( \chi^2 \)-conditional distribution of \( R \) as derived in Loftus & Taylor (2015) for (3) when \( \rho = 0 \) no longer holds for more general hypotheses.

Similar to (2), Yang et al. (2016) decompose \( \mathbf{Y} \) as \( R \cdot \text{dir}_{\mathbf{W}}(\mathbf{y}) + \mathbf{P}^\perp_{\mathbf{W}} \mathbf{Y} \) and condition on \( \text{dir}_{\mathbf{W}}(\mathbf{Y}) = \text{dir}_{\mathbf{W}}(\mathbf{y}) \) as well as on \( \mathbf{P}^\perp_{\mathbf{W}} \mathbf{Y} = \mathbf{P}^\perp_{\mathbf{W}} \mathbf{y} \). Then, the only variation left is in \( R \) and the selection \( \mathcal{A} \) can be equally written as \( R \in \mathcal{R}_y \) with

\[
\mathcal{R}_y = \left\{ R > 0 : \mathcal{S}(R \cdot \text{dir}_{\mathbf{W}}(\mathbf{y}) + \mathbf{P}^\perp_{\mathbf{W}} \mathbf{y}) = \mathcal{A} \right\}.
\]

Yang et al. (2016) then derive the conditional distribution of \( R \), conditional on \( \text{dir}_{\mathbf{W}}(\mathbf{y}) \) as well as on \( \mathbf{P}^\perp_{\mathbf{W}} \mathbf{y} \). The corresponding density is

\[
f(R) \propto R^{w-1} \exp\left\{-\frac{1}{2\sigma^2}(R^2 - 2R \cdot \langle \text{dir}_{\mathbf{W}}(\mathbf{y}), \mathbf{\mu} \rangle)\right\} \cdot \mathbb{1}\{R \in \mathcal{R}_y\}
\]

with indicator function \( \mathbb{1}\{\cdot\} \). (5) can be used to conduct inference on the inner product \( \langle \text{dir}_{\mathbf{W}}(\mathbf{y}), \mathbf{\mu} \rangle \). As for the quantity of interest \( \rho = ||\mathbf{P}_\mathbf{W}\mathbf{\mu}||_2 \geq \langle \text{dir}_{\mathbf{W}}(\mathbf{y}), \mathbf{\mu} \rangle \) holds, (5) can also be used to construct a lower bound for \( \rho \).

3.4 Inference without explicit inference region definition

Whereas most approaches for selective inference require an explicit definition of the space \( \mathcal{G} \), to which \( \mathbf{Y} \) is restricted by the selection procedure, a mathematical description of \( \mathcal{G} \) is not always feasible. However, as pointed out by Fithian et al. (2014), Yang et al. (2016), such a characterization is not mandatory when sampling from the conditional distribution of \( \mathbf{Y} \) is possible. In the following, we describe the idea of Yang et al. (2016), who use an importance sampler when conducting inference for (3).

Theorem 1 in Yang et al. (2016) states that, conditional on \( \text{dir}_{\mathbf{W}}(\mathbf{y}) \), \( \mathbf{P}^\perp_{\mathbf{W}} \mathbf{y} \) and the
selection event, inference can be conducted using

$$\zeta(t) = \frac{\int_{R \in \mathcal{R}_y, R > \|P_W y\|_2} R^{w-1} e^{-(R^2 - 2 R t)/2 \sigma^2} dR}{\int_{R \in \mathcal{R}_y} R^{w-1} e^{-(R^2 - 2 R t)/2 \sigma^2} dR}$$

as $$\zeta(t_y)$$, a p-value for $$H_0 : \langle \text{dir}_W(y), \mu \rangle = t_y$$, is Uniform[0, 1]-distributed. Here, $$\zeta(\cdot)$$ can also be seen as the survival function derived from the density defined in (5). In order to circumvent an explicit definition of the selection region $$\mathcal{R}_y$$, the authors note that (6) is equal to

$$\mathbb{E}_{R \sim \sigma \chi_w}(e^{R t / \sigma^2} \cdot 1\{R \in \mathcal{R}_y, R > \|P_W y\|_2\}) \mathbb{E}_{R \sim \sigma \chi_w}(e^{R t / \sigma^2} \cdot 1\{R \in \mathcal{R}_y\}).$$

which can be approximated by the ratio of empirical expectations computed with a large number of samples $$r^b \sim \sigma \cdot \chi_w, b = 1, \ldots, B$$. In particular, to evaluate the argument of both expectations in (7) for some $$r^b$$, $$r^b \in \mathcal{R}_y$$ must be checked. To this end, note that the only variation of $$(Y \mid \text{dir}_W(y), P_W y)$$ is in $$R$$. We therefore define $$y^b = P_W y + r^b \cdot \text{dir}_W(y)$$ and rerun the algorithm to check whether $$S(y^b) = A$$, or equivalently, whether $$r^b \in \mathcal{R}_y$$.

Drawing samples from the $$\sigma \chi_w$$-distribution, however, is less promising when $$\|P_W y\|_2$$ is large. In this case, $$\mathbb{P}(R \in \mathcal{R}_y)$$ may be very small and an excessively large number of samples is needed to obtain a good approximation of $$\zeta(t)$$. Yang et al. (2016) therefore suggest an importance sampling algorithm, which draws new samples $$r^b$$ from a proposal distribution $$F_{\text{prop}}$$ such as $$\mathcal{N}(\|P_W y\|_2, \sigma^2)$$ with density $$f_{\text{prop}}$$ and then approximates (7) by

$$\zeta(t) \approx \zeta(t) = \frac{\sum_b w_b \cdot e^{r^b t / \sigma^2} \cdot 1\{r^b \in \mathcal{R}_Y, r^b > \|P_W y\|_2\}}{\sum_b w_b \cdot e^{r^b t / \sigma^2} \cdot 1\{r^b \in \mathcal{R}_Y\}}$$

with sampling weights $$w_b = f_{\sigma \chi_w}(r^b) / f_{\text{prop}}(r^b)$$. 

4 Selective Inference concepts for $$L_2$$-Boosting

We now present selective inference concepts for $$L_2$$-Boosting. In Section 4.1 we first show how to use inference concepts proposed for the Lasso and forward stagewise algorithms by deriving a polyhedron representation of selection conditions in $$L_2$$-Boosting. After evaluating the resulting inference framework in Section 4.2, we propose an alternative concept for $$L_2$$-Boosting and similar learners (slow learners), which can repeatedly select the same base-learner and change estimated regression coefficients only incrementally. Based on this
idea, we derive a powerful inference framework for \( L_2 \)-Boosting with linear base-learners in Section 4.3 and describe important extensions in Section 4.4.

4.1 Polyhedron representation-based inference for \( L_2 \)-Boosting

Consider using \( L_2 \)-Boosting with only linear base-learners, i.e., \( D_j = 0, \text{ncol}(X_j) = 1 \forall j \), to fit a linear regression model. Following Tibshirani et al. (2016) we can derive a polyhedron representation \( G = \{ y : \Gamma y \geq b \} \) in a similar fashion to other stepwise regression procedures for the given selection path \( j^{(1)}, \ldots, j^{(m_{\text{stop}})} \) of \( L_2 \)-Boosting.

This can easily be proven by regarding the residual vector \( u^{(m)} \) of step \( m \) as a function of \( y \). The selection condition for the \( m \)th chosen base-learner

\[
|| (I - H_{j^{(m)}})u^{(m)} ||^2 \leq ||(I - H_j)u^{(m)}||^2 \quad \forall j \neq j^{(m)}
\]

\[
\Leftrightarrow \left( s_m X_{j^{(m)}}/(||X_{j^{(m)}}||_2 \pm X_j^\top/||X_j||_2) \right) u^{(m)} \geq 0 \quad \forall j \neq j^{(m)}.
\]

with \( s_m = \text{sign}(X_{j^{(m)}}^\top u^{(m)}) \), can be written as affine restriction on \( y \) by plugging

\[
u^{(m)} = \left[ \prod_{l=1}^{m-1} (I - \nu H_{j^{(m-l)}}) \right] y =: \Upsilon^{(m)} y
\]

into (9). For a given selection path and list of signs \( s_m, m = 1, \ldots, m_{\text{stop}} \) this yields the polyhedron representation \( G \) with fixed \( (2 \cdot (p - 1) \cdot m_{\text{stop}}) \times n \) matrix \( \Gamma \) as stacked matrix of \( n \)-dimensional row vectors, where the rows \( \Gamma_{(\tilde{m}+2(j-\omega(j))-1):\tilde{m}+2(j-\omega(j))} \) with \( \tilde{m} = 2 \cdot (p - 1) \cdot (m - 1) \) and \( \omega(j) = 1 \{ j > j^{(m)} \} \) are given by

\[
(s_m X_{j^{(m)}}/(||X_{j^{(m)}}||_2 \pm X_j^\top/||X_j||_2) \Upsilon^{(m)}) \quad \forall j \neq j^{(m)}.
\]

As for other procedures described in the post-selection inference literature, this representation only holds if the columns of \( X \) are in general position, which however, is not a very stringent assumption (see, e.g., Tibshirani et al. 2016, section 4).

By showing that the \( L_2 \)-Boosting path results in a space restriction for \( Y \), which can be described as a polyhedral set, conditional on the list of signs, quantities of interest \( v^\top \mu \) can be tested based on the conditional distribution of \( v^\top Y | Y \in G \) as proposed by Tibshirani et al. (2016). To this end, we have to condition on the selection path. If we do not additionally condition on the list of signs, \( G \) is a union of polyhedra (cf. Lee et al.)
Similar, for analytic space restrictions other than polyhedra, e.g., when using group base-learners or base-learners with penalties, inference can be conducted for $L_2$-Boosting following the work of [Loftus & Taylor (2015), Rügamer & Greven (2018)].

4.2 Choice of the Conditioning Event for Slow Learners

For the selection approaches discussed in Section 3, conditioning on the selection path is equivalent to conditioning on the selected model, which helps in deriving the corresponding conditional distribution. For boosting and other slow learners that can repeatedly select the same base-learner, conditioning on the selection path and thus on variable selection decisions in each algorithmic step will result in a loss of power. In fact, such a conditional inference will have almost no power in most practically relevant situations, as we show empirically for the polyhedron approach in the simulation section. In order to avoid excessive conditioning, we propose to condition only on the set of selected covariates, i.e., on the selected statistical model.

Conditioning only on the selected covariates, however, means that the mathematical description of the inference region becomes far more difficult. For $L_2$-Boosting with linear base-learners, this would result in a union of not necessarily overlapping polyhedra for the different selection paths leading to the same selected model. In particular for $L_2$-Boosting, we do not think that a general analytical description of the inference region is possible. We thus circumvent this problem using a Monte Carlo approximation, adapting and extending the existing approaches presented in Section 3.

4.3 Powerful Inference for $L_2$-Boosting with Linear Base-learners

We now build on the ideas of Section 3.2 and 3.4 to practically realize the idea of the previous Section 4.2. We base inference on the potentially multiply truncated Gaussian distribution of $R = v^\top Y$ conditional on $P_v \perp y$ and the selection $R \in \mathcal{R}_y$. Then, the truncated normal density of $R$ is given by

$$f(R) \propto \exp \left\{ -\frac{1}{2\sigma^2 v^\top v} (R - v^\top \mu)^2 \right\} \cdot 1(R \in \mathcal{R}_y),$$

(11)
where $\mathcal{R}_y$ is a union of polyhedra. The proof of equation (11) follows analogously to Lemma 1 of Yang et al. (2016) for $R = v^\top Y$ using $w = 1$ and $\langle \text{dir}_W(y), \mu \rangle = v^\top \mu / ||v||_2$ in this case (with a rescaled definition of $R$ and keeping the sign by using a normal instead of a $\chi$-distribution). Let $r_{\text{obs}} = v^\top y$. Then, analogous to Yang et al. (2016) we can define a p-value by

$$P = \frac{\int_{R > r_{\text{obs}}, R \in \mathcal{R}_y} e^{-(2\sigma^2 v^\top v)^{-1} R^2} \, dR}{\int_{R \in \mathcal{R}_y} e^{-(2\sigma^2 v^\top v)^{-1} R^2} \, dR}$$

for $H_0 : v^\top \mu = 0$ and since the truncated Gaussian distribution with potentially multiple truncation limits is monotonously increasing in its mean $\rho$ (see, e.g., Rügamer & Greven 2018), we can find unique values $\rho_{\alpha/2}, \rho_{1-\alpha/2}$ for any $\alpha \in (0, 1)$, such that

$$\varsigma(\rho_a) = \frac{\int_{R > r_{\text{obs}}, R \in \mathcal{R}_y} e^{-(2\sigma^2 v^\top v)^{-1} (R^2 - 2R\rho_a)} \, dR}{\int_{R \in \mathcal{R}_y} e^{-(2\sigma^2 v^\top v)^{-1} (R^2 - 2R\rho_a)} \, dR} = a, \quad a \in \{\alpha/2, 1 - \alpha/2\}$$

to construct a two-sided confidence interval $[\rho_{\alpha/2}, \rho_{1-\alpha/2}]$.

Note that $P = \varsigma(0)$, and $\varsigma(\rho_a)$ can then be rewritten as

$$\frac{\mathbb{E}_{R \sim \mathcal{N}(0, \sigma^2 v^\top v)} \left[ \mathbb{1}\{R \in \mathcal{R}_y, R > r_{\text{obs}}\} \cdot e^{(\sigma^2 v^\top v)^{-1} R\rho_a} \right]}{\mathbb{E}_{R \sim \mathcal{N}(0, \sigma^2 v^\top v)} \left[ \mathbb{1}\{R \in \mathcal{R}_y\} \cdot e^{(\sigma^2 v^\top v)^{-1} R\rho_a} \right]},$$

which allows for an empirical approximation as in (8). Further note that this approach does not require to condition on the list of signs and therefore can also be used to compute inference for, e.g., a certain selection event of the Lasso without conditioning on the corresponding list of signs.

**Monte Carlo Approximation**

In practice, importance sampling from $\Pi = \mathcal{N}(r_{\text{obs}}, \sigma^2 v^\top v)$ works well if truncation limits around $r_{\text{obs}}$ are fairly symmetric, yielding the weights $w_b = \exp((2r^b r_{\text{obs}} - r_{\text{obs}}^2)/(-2\sigma^2 v^\top v))$ for the importance sampler. A refinement of the sampling routine is necessary to also work well in more extreme cases. An example frequently encountered in practice is given when $r_{\text{obs}}$ is rather large and at the same time lies very close to one truncation limit, yielding an insufficient number of samples $r^b \in \mathcal{R}_y$ to approximate the truncated distribution well.

We therefore propose a more efficient sampling routine, motivated by and applicable to
selection procedures, for which the support of the truncated distribution is known to be a single interval \([V_{\text{lo}}, V_{\text{up}}]\). Our idea is that, in this case, we do not need to characterize the space empirically since the distribution of interest is known with the exception of the interval limits (the variance is assumed to be known and the null distribution determines the mean \(\rho\)). By employing a line search, we can find \([V_{\text{lo}}, V_{\text{up}}]\) and conduct inference based on the truncated normal distribution function \(F_{\rho,\sigma^2v^\top v}(\cdot)\). We use such a corresponding line search here to refine the importance sampling. By searching through the space of potential values \(R \in \mathcal{R}_y\), a preliminary interval \([\tilde{R}_{\text{lo}}, \tilde{R}_{\text{up}}]\) covering \(\mathcal{R}_y\) can be found with typically negligible computational cost. By, e.g., successively checking extreme quantiles of \(\Pi\) for their congruency with respect to \(\mathcal{R}_y\) using on the order of 50 refits of the model and defining \(\tilde{R}_{\text{lo}}, \tilde{R}_{\text{up}}\) such that both limits include all values, for which \(R \in \mathcal{R}_y\), we can find a superset of the support of \(R\) up to numerical precision. We then draw from a uniform distribution with support \([\tilde{R}_{\text{lo}}, \tilde{R}_{\text{up}}]\). In comparison to simply draws sampling from \(\Pi\), finding preliminary truncation limits \([\tilde{R}_{\text{lo}}, \tilde{R}_{\text{up}}]\) to refine the sampling space prior to the actual sampling proves to notably enhance accuracy and efficiency due the increased number of accepted samples.

### 4.4 Further extensions

The ideas of section 4.2 and 4.3 can be extended to allow for computations in further relevant settings. We additionally discuss four practically important extensions in the following.

**Inference for groups of variables.** In order to test groups of variables, the approach by Yang et al. (2016) described in Subsection 3.4 can almost directly be applied. To this end, we define \(S\) based on the set of chosen variables and use the sampling approach proposed in Subsection 4.3 for the \(\chi\)-distribution on \(\mathbb{R}^+\), such that \(\tilde{R}_{\text{lo}} \geq 0\).

**Incorporating cross-validation and other sub-sampling techniques.** One of the most common ways to choose a final stopping iteration for the boosting algorithm is by using a resampling technique such as \(k\)-fold cross-validation (CV) and estimating the prediction error of the model in each step. By choosing the model with the smallest estimated prediction error,
we again exploit information from the data, which we have to discard in the following inference. $S$ then corresponds to the selection obtained using $L_2$-Boosting with stopping iteration chosen by CV. We can extend the sampling approach described in Section 4.3 by incorporating the CV conditions into the space definition of $R_y$. Therefore, define a (multivariate) random variable $\Delta$ describing these conditions, which is independent of $Y$. For $k$-fold CV, for example, $\Delta$ is a uniformly distributed random variable on all possible permutations of $(1, \ldots, 1, 2, \ldots, 2, \ldots, k, \ldots, k)$ yielding the assignments $\delta = (\delta_1, \ldots, \delta_n)$ for every entry in $y$ to one of the $k$-folds with equal probability (if $n$ is a multiple of $k$). For conducting inference, we additionally condition on $\Delta = \delta$, i.e., we keep the folds fixed and identical to those for the original fit, when rerunning the algorithm with a new sample $y^b$ to check for consistency with the observed selection event $R_y$. In fact, this approach is not only restricted to resampling methods. Stability selection (Shah & Samworth 2013) or other possibilities to choose an “optimal” number of iterations, as for example, by selection criteria such as the Akaike Information Criterion (AIC, Akaike 1974) can be incorporated into the inference framework in the same manner. For a mathematical justification observe that conditional on the selection event $R_y$ (including conditions on other random variables such as $\Delta$), $P_W$ is fixed and Lemma 1 by Yang et al. (2016) holds analogously.

*Unknown error variance.* If the true error variance is unknown, we may use a consistent estimator instead. Judging by our simulation results, the effect of plugging in the empirical variance of the boosting model residuals is negligible in many cases and may also be a better (less anti-conservative) choice than the analogous estimator given by ordinary least squares estimation in the selected model due to the shrinkage effect. In cases with smaller signal-to-noise ratio, however, the plug-in approach may also yield invalid p-values under the null as shown in our simulation section. Tibshirani et al. (2018) present a plug-in as well as a bootstrap version of the test statistic, which yield asymptotically conservative p-values for $v^\top \mu = 0$. The bootstrap approach, however, can only be conducted efficiently if truncation limits of the test statistic are known. In the simulation section, we investigate the first suggestion by Tibshirani et al. (2018) – using the empirical variance of $y$ as a conservative estimate for $\sigma^2$ – which better suits the presented framework.
Smooth effects. The given approach can also be used for additive models when the linear predictor \( \eta_i = x_i^\top \beta \) in the working model \( y_i = \eta_i + \varepsilon_i, i = 1, \ldots, n \) is extended by additive terms of the form \( g(c_i) \) for some covariate \( c = (c_1, \ldots, c_n)^\top \). For the ease of presentation, we assume only one covariate \( c \) that is incorporated with an additive term, but the general case is analogous. We use a basis representation \( g(c_i) = B(c_i)\gamma = \sum_{\omega=1}^{M} B_{\omega}(c_i)\gamma_{\omega} \) with \( M \) basis function \( B_{\omega}(\cdot) \) evaluated at the observed value \( c_i \), basis coefficients \( \gamma_{\omega}, B(c_i) = (B_1(c_i), \ldots, B_M(c_i)) \) and \( \gamma = (\gamma_1, \ldots, \gamma_M)^\top \). When \( X_A \) is the composed matrix of all covariates, which are assumed to have a linear effect, and of the evaluated basis functions \( \tilde{B} = (B(c_1)^\top, \ldots, B(c_n)^\top)^\top \) of \( c \), we again might be interested in testing the best linear approximation of \( \mu \) in the space spanned by a given design matrix \( X_A \). To this end, we can perform a point-wise test \( H_0 : g(c) = 0 \) for some \( c \), where \( g \) is the “true” function in the basis space resulting from the best linear approximation of \( \mu \) by the given model. This can be done by using the proposed framework with test vector \( v^\top = B^0(c)(X_A^\top X_A)^{-1}X_A^\top \), as \( g(c) = v^\top \mu \), where \( B^0(c) \) has the same structure as one row of \( X_A \) but with all columns except those corresponding to \( B(c) \) set to zero. Instead of a point-wise test, the whole function can be tested

\[ H_0 : g(\cdot) \equiv 0 \]  

by regarding the columns in \( \tilde{B} \) as groups of variables and setting \( W \) in (3) to \( P_{X_A}^\perp \tilde{B} \), where \( X_A \setminus j \) denotes \( X_A \) without the \( p_j \) columns of \( \tilde{B} \).

The proposed tests and testvectors \( v \) or matrices \( W \) can also be used when smooth effects are estimated using a penalized base-learner with \( D_j \neq 0 \). We note that this is one of the advantages of \( L_2 \)-Boosting over the Lasso, as fitting smooth effects is not straightforward for the Lasso.

5 Simulations

We now provide evidence for the validity of our method for linear and spline base-learners based on \( B = 1000 \) samples per iteration and \( \varrho = 1000 \) simulation iterations. We also show the performance of the proposed method in comparison to the polyhedron approach in a
relevant setting and investigate the effect of different variance values. For linear regression with linear base-learners the true underlying model is given by

\[ y_i = \eta_i + \varepsilon_i = X[i,1:4] \beta + \varepsilon_i, \quad i = 1, \ldots, n, \]  

(13)

where \( \beta = (4, -3, 2, -1)^T \), \( \eta = (\eta_1, \ldots, \eta_n)^T \), \( \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \) with \( \sigma \) defined such that the signal-to-noise ratio \( \text{SNR} := \frac{\text{sd}(\eta)}{\sigma} \in \{1, 4\} \) and \([i, 1:4]\) indicates the rows and columns of \( X \), respectively. We construct four linear base-learners for the four covariates \( x_1, \ldots, x_4 \) in \( X[i,1:4] \) and additionally build \( p_0 \in \{4, 22\} \) base-learners based on noise variables for \( n \in \{25, 100\} \) observations, where the columns in \( X \) are independently drawn from a standard normal distribution (empirical correlations range from \(-0.53\) to \(0.48\)). Note that the case \( p_0 = 22 \) and \( n = 25 \) therefore also includes a setting, in which \( p > n \) holds.

Figure 1 shows the observed p-values versus the expected quantiles of the standard uniform distribution for settings, in which either the true model or a model larger than the true model with all four signal variables is selected. This corresponds to selection events, in which the null hypothesis (1) holds for \( j > 4 \) and thus p-values of inactive variables should exhibit uniformity given the selection event \( A \). The mixture of uniform \( U[0,1] \) p-values when aggregating across selected models again results in \( U[0,1] \) p-values. Results are given in Figure 1 (\( n = 25 \)) and in Figure 2 in the Online Appendix (\( n = 100 \)).

Results: p-values for effects of “true effect” variables show deviations from the angle bisecting line, indicating the ability of the proposed procedure to correctly infer the significance of the effects. The power decreases for a smaller number of observations (cf. Figure 2), a smaller SNR and a larger number of noise variables. The polyhedron approach yields correct p-values under the null, but shows no power for non-noise variables. p-values for the proposed approach are uniform under the null when using the true variance (even when selecting \( m_{\text{stop}} \) using CV), with more conservative results when using the empirical variance of the response and slightly non-uniform p-values when using a plugin estimator. Differences are similar for larger \( n \). In this respect, the empirical variance of boosting residuals is more favorable than that of an OLS refit, but can also lead to deviations. However, note that the empirical approximation of p-values is not very accurate in the settings where specific selection events are rather unlikely, as only a small number of samples \( r^b \in \mathcal{R}_y \) can be used.
Figure 1: Observed p-values vs. expected quantiles across different covariates (rows) as well as different methods, number of noise variables, number of boosting iterations and SNR (columns) after boosting with a step-length of 0.1 using different variance types (colors), B = 1000, and a total of 1000 simulation iterations in settings with n = 25. p-values are shown for simulation iterations, in which either the true model or a model larger than the true model is selected. For each setting, the number of retained iterations (nobs) is noted in the left upper corner.

Corresponding confidence intervals of the proposed test procedure reveal approximately (1−α)% coverage for the same simulation settings. Results for α = 0.05 are given in Table 1. Deviations from the ideal coverage of 95% are primarily due to numerical imprecision when inverting the hypothesis test and more accurate results can be obtained in applications when the number of non-rejected samples is too low by simply increasing the number of samples B.

In the Supplementary Material, we additionally provide results for other settings of the
Table 1: Estimated coverage of selective confidence intervals obtained by the proposed sampling approach for \( n = 25 \) observations when using the true variance in different settings (columns) in which either the true model or a model larger than the true model is selected.

| \( p_0 \), number of iterations, SNR | 4, 40, 1 | 4, 80, 1 | 4, CV, 1 | 22, 40, 1 | 22, 40, 4 |
|-------------------------------------|---------|---------|---------|-----------|-----------|
| coverage noise variables            | 0.9566  | 0.9571  | 0.9618  | 0.9485    | 0.9211    |
| coverage signal variables           | 0.9699  | 0.9559  | 0.9326  | 0.9444    | 0.9429    |

previous simulation study as well as results for additive models using spline base-learners, where the true underlying function is given by

\[
y_i = \sin(2X_{[i,1]}) + \frac{1}{2}X_{[i,2]}^2 + \varepsilon_i, \quad i = 1, \ldots, n = 300, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)
\]

with \( \sigma \) defined such that the signal-to-noise ratio \( \text{SNR} = 0.5 \) and 13 further covariates \( X_{[3:15]} \). All covariate effects are represented using penalized B-splines (P-spline; [Eilers & Marx 1996]) with B-Spline basis of degree 3, 5 knots and second order differences penalty. Tests for the whole function are performed as proposed in [12]. Results suggest very high power but uniformity of p-values for noise variables, supporting the conclusion that the proposed test also works well for additive terms. We further compare the selective approach for linear base-learners with the naive approach, thereby illustrating the invalidity of classical unadjusted inference (see Figure 2), compare the length of selective and naive intervals (Figure 3) and address the criticism of potentially infinite selective intervals by investigating the frequency of an infinite interval for two simulation scenarios for \( n = 100 \) and \( p = 26 \) (Figure 4). In this case, infinite length of corresponding intervals occurs only in around 5% of all cases.

5.1 Computation time and further details

As the proposed framework requires refitting the selection procedure \( B \) times, the computation time might be the biggest concern for practitioners. When it is not possible to parallelize the model fits for the values \( r^b \), increasing \( B \) obviously results in a linear increase of computation similar to conducting a bootstrap. In comparison to the model refits, the preceding line search for the limits of \( \mathcal{R}_Y \) can be rather cheap, but may take a predominant amount of time for very rare events. For these rare events, practitioners have the
choice to either avoid extended run-times by using a sampling approach without a preceding search for the limits of $R_Y$ or to obtain more accurate inference results by using the line search approach with additional run-time. We note that without preceding line search, however, sampling may yield a very small number of un-rejected samples in this case and lead to a higher inaccuracy in inference statements. In order to give a rough insight into run-times for our software, we provide computation times for the sampling itself using different settings of $n$ and $p$ which include realistic, high-dimensional setups after model selection with subsequent 5-fold CV. Estimated run-times with parallelization of the 5-fold CV but without parallelization of the refitting procedure itself are shown in Figure 7 in the Online Appendix D for inference statements on one hypothesis (one projection direction) based on 5 replications per setting. Results suggest that computation time is sublinear in $n$, which is due to the fact, that the hat matrix will only be computed once for all refits, but approximately $O(p^2 \log(p))$.

6 Application

We now apply our framework to the prostate cancer data set (Stamey et al. 1989) to model logarithmic PSA level ($lpsa$) of patients having prostate cancer. This data set has already been analyzed with regard to post-selection inference by, for example, Tibshirani et al. (2016) using forward stepwise regression and testing after a prespecified number of steps. In contrast to previous approaches, we can choose the stopping iteration using CV and do not need to enforce effects of continuous covariates to be linear. Instead, we assume a more flexible additive model

$$lpsa_i = \beta_0 + \sum_{j=1}^{7} g_j(X_{i,j}) + \sum_{j=2}^{4} I(\text{gleason}_i = j)\beta_j + \varepsilon_i, \quad i = 1, \ldots, 97,$$

with 7 metric variables $X_{j}, j = 1, \ldots, 7$ and categorical variable $\text{gleason}$ that can check the linearity assumption previously imposed. In order to estimate the smooth effects, we fit the model using cubic P-spline base-learners with second-order difference penalties. To facilitate a fair base-learner selection (Hofner et al. 2011), we split up effects of continuous covariates into a linear effect and a non-linear deviation from the corresponding linear effect and penalize the categorical variable using a Ridge penalty. The optimal stopping
iteration \( m_{\text{stop}} = 47 \) for the boosting algorithm with step-length \( \nu = 0.1 \) is found by using 10-fold CV, which is incorporated into the selection mechanism \( S \). After 47 iterations, five effects are selected by the boosting procedure, including two non-linear deviations for the covariate \( \text{lbph} \) (logarithmic benign prostatic hyperplasia amount) and the covariate \( \text{pgg45} \) (percentage Gleason scores 4 or 5). The two covariates show an inverse U-shaped effect, which is shown in the Online Appendix. Table 2 shows the results for component-wise tests of linear and additive terms for hypothesis tests based on the proposed sampling approach with \( B = 5000 \) samples. Testing additive terms, which have been split up into a linear part and a non-linear deviation, can be done by defining \( \tilde{B} \) as concatenated matrix of the covariate vector itself and the corresponding matrix of evaluated basis functions orthogonalized to the linear effect. The logarithmic cancer volume (\( \text{lvacol} \)) is found to be the only variable having a significant influence on the response for the given model.

Table 2: Magnitude and sign of linear (L) and magnitude of non-linear (NL) projections \( \|P_W y\|_2 \) for the selected model terms \( \text{lbph} \) (logarithmic benign prostatic hyperplasia amount), \( \text{pgg45} \) (percentage Gleason scores 4 or 5), \( \text{lcavol} \) (logarithmic cancer volume), \( \text{lweight} \) (logarithmic prostate weight) and \( \text{svi} \) (seminal vesicle invasion) as well as corresponding lower confidence interval limits and p-values. In the linear case the magnitude corresponds to the absolute value of the OLS effect estimate.

|         | lbph (NL) | pgg45 (NL) | lcavol (L) | lweight (L) | svi (L) |
|---------|-----------|------------|------------|-------------|--------|
| magnitude (sign) | 2.3319    | 2.8518     | 4.0992 (+) | 2.2067 (+)  | 1.9520 (+) |
| lower limit | 0         | 0          | 2.4859     | 0           | 0      |
| p-value    | 0.3452    | 0.2467     | 0.0004     | 0.3752      | 0.1212 |

7 Discussion

In this paper we review several recently proposed selective inference frameworks and transfer and adapt them to the \( L_2 \)-Boosting algorithm. As far as we know, there are no previous general methods available to quantify uncertainty of boosting estimates (or more generally for slow learners) in a classical statistical manner when variable selection is performed. We propose tests and confidence intervals for linear base-learners as well as for group vari-
able and penalized base-learners. Using Monte Carlo approximation for the calculation of p-values and confidence intervals, we avoid the necessity for an explicit mathematical description of the inference space. This allows us to condition on less, which in turn increases power notably in comparison to Polyhedron approaches. We apply our framework to the prostate cancer data set and in contrast to published analyses of this data also allow for non-linear partial effects as well as selection of the stopping iteration using CV. Using simulation studies with a range of settings, we verify the properties of our approach.

This work opens up a variety of future research topics. In order to leave more information for inference and further reduce the occurrence of infinite confidence intervals, the framework could, e.g., be extended by incorporating randomization in the model selection and inference step (see, e.g. Tian Harris et al. 2016). Adapting this concept for the given framework is, however, not straightforward as it is not clear, whether estimators obtained by the boosting procedure are the solution to a closed-form optimization problem.

An extension to generalized linear models (GLMs) and beyond also proves to be difficult since conditions involving $y$ might imply conditioning on $y$ itself if the response is discrete (see Fithian et al. 2014 for more details on selective inference for GLMs). It would also be interesting to investigate whether asymptotic results of Tian & Taylor (2017) can be used to construct inference for CFGD algorithms other than $L_2$-Boosting.
References

Akaike, H. (1974), ‘A new look at the statistical model identification’, *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L. et al. (2013), ‘Valid post-selection inference’, *The Annals of Statistics* **41**(2), 802–837.

Brockhaus, S., Fuest, A., Mayr, A. & Greven, S. (2018), ‘Signal regression models for location, scale and shape with an application to stock returns’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(3), 665–686.

Brockhaus, S., Scheipl, F., Hothorn, T. & Greven, S. (2015), ‘The functional linear array model’, *Statistical Modelling* **15**(3), 279–300.

Bühlmann, P. & Hothorn, T. (2007), ‘Boosting algorithms: Regularization, prediction and model fitting (with discussion)’, *Statistical Science* **22**(4), 477–505.

Bühlmann, P. & Yu, B. (2003), ‘Boosting with the $L_2$ loss: regression and classification’, *Journal of the American Statistical Association* **98**(462), 324–339.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004), ‘Least angle regression’, *The Annals of Statistics* **32**(2), 407–499.

Eilers, P. H. C. & Marx, B. D. (1996), ‘Flexible smoothing with B-splines and penalties’, *Statistical Science* **11**(2), 89–121.

Fithian, W., Sun, D. & Taylor, J. (2014), ‘Optimal Inference After Model Selection’, *arXiv e-prints arXiv:1410.2597* .

Friedman, J. H. (2001), ‘Greedy function approximation: A gradient boosting machine’, *The Annals of Statistics* **29**(5), 1189–1232.

Hofner, B., Hothorn, T., Kneib, T. & Schmid, M. (2011), ‘A framework for unbiased model selection based on boosting’, *Journal of Computational and Graphical Statistics* **20**(4), 956–971.
Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. & Hofner, B. (2010), ‘Model-based boosting 2.0’, *Journal of Machine Learning Research* **11**(Aug), 2109–2113.

Kivaranovic, D. & Leeb, H. (2018), ‘Expected length of post-model-selection confidence intervals conditional on polyhedral constraints’, *ArXiv e-prints*.

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016), ‘Exact post-selection inference, with application to the lasso’, *The Annals of Statistics* **44**(3), 907–927.

Loftus, J. R. & Taylor, J. E. (2014), ‘A significance test for forward stepwise model selection’, *arXiv e-prints arXiv:1405.3920*.

Loftus, J. R. & Taylor, J. E. (2015), ‘Selective inference in regression models with groups of variables’, *arXiv e-prints arXiv:1511.01478*.

Luo, Y. & Spindler, M. (2017), ‘L2-boosting for economic applications’, *American Economic Review* **107**(5), 270–273.

Mayr, A., Hofner, B., Waldmann, E., Hepp, T., Meyer, S. & Gefeller, O. (2017), ‘An update on statistical boosting in biomedicine’, *Computational and Mathematical Methods in Medicine* **2017**, 12.

Melcher, M., Scharl, T., Luchner, M., Striedner, G. & Leisch, F. (2017), ‘Boosted structured additive regression for escherichia coli fed-batch fermentation modeling’, *Biotechnology and Bioengineering* **114**(2), 321–334.

Rügamer, D., Brockhaus, S., Gentsch, K., Scherer, K. & Greven, S. (2018), ‘Boosting factorspecific functional historical models for the detection of synchronization in bioelectrical signals’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(3), 621–642.

Rügamer, D. & Greven, S. (2018), ‘Selective inference after likelihood- or test-based model selection in linear models’, *Statistics & Probability Letters* **140**, 7 – 12.

Shah, R. D. & Samworth, R. J. (2013), ‘Variable selection with error control: Another look at stability selection’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(1), 55–80.
Stamey, T. A., Kabalin, J. N., Mcneal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A. & Yang, N. (1989), ‘Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients’, *The Journal of Urology* **141**(5), 1076 – 1083.

Tian Harris, X., Panigrahi, S., Markovic, J., Bi, N. & Taylor, J. (2016), ‘Selective sampling after solving a convex problem’, *ArXiv e-prints*.

Tian, X. & Taylor, J. (2017), ‘Asymptotics of selective inference’, *Scandinavian Journal of Statistics* **44**(2), 480–499.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R. & Wasserman, L. (2018), ‘Uniform asymptotic inference and the bootstrap after model selection’, *The Annals of Statistics* **46**(3), 1255–1287.

Tibshirani, R. J., Taylor, J., Lockhart, R. & Tibshirani, R. (2016), ‘Exact post-selection inference for sequential regression procedures’, *Journal of the American Statistical Association* **111**(514), 600–620.

Wasserman, L. & Roeder, K. (2009), ‘High dimensional variable selection’, *The Annals of Statistics* **37**(5A), 2178–2201.

Yang, F., Barber, R. F., Jain, P. & Lafferty, J. (2016), Selective inference for group-sparse linear models, in ‘Advances in Neural Information Processing Systems’, pp. 2469–2477.

Yuan, M. & Lin, Y. (2006), ‘Model selection and estimation in regression with grouped variables’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.
Appendix

Appendix A: Further Simulation Results

A.1 Further Simulation Results for Linear Base-learners

We first investigate the validity of our inference approach in two additional settings for \( n = 100 \) observations. The results are visualized in Figure 2, suggesting powerful and valid inference if the selective approach is used and proving the invalidity of classical inference (naive) when not adjusted for model selection.

![Figure 2: Observed p-values vs. expected quantiles across different covariates (columns) as well as different SNR (rows) for boosting with different variance values / estimates (colours), 26 variables including 22 noise variables, \( B = 1000 \), a total of 100 simulation iterations and \( n = 100 \) (in contrast to \( n = 25 \) in the main article). P-values are shown for simulation iterations, in which either the true model or a model larger than the true model is selected. For each setting, the number of iterations (nobs) is noted in the left upper corner.

We further use the simulation scenario used for Figure 2 to examine the length of selective confidence intervals in comparison to naive confidence intervals (Figure 3) and investigate the frequency of observing an infinite length due to one or two infinite interval limits (Figure 4). Note that the given frequencies in Figure 4 are an upper bound approximation since infinite interval limits can also occur due to the Monte Carlo approach if not enough samples are congruent with the initial selection.
Figure 3: Ratio of selective confidence interval length divided by the classical confidence interval length for different SNR (rows) and variances (colours) used for the computation of the distribution of test statistic. Note that the y-axis is on a logarithmic scale.

A.2 Further Simulation Results for P-spline Base-learners
Figure 4: Frequency of finite / infinite interval lengths in two SNR settings (columns) for 100 simulation iterations. Iterations, for which the corresponding variable was not selected, do not contribute to the bars. Variables 5 - 26 correspond to noise variables.

Figure 5: Observed p-values vs. expected quantiles across different covariates (rows) as well as different variance values / estimates (colours) for two different SNR (colours) for testing a function using boosted P-spline baselearners after 50 iterations and a step-length of 0.1, using a total of 500 simulation iterations. p-values are shown for simulation iterations, in which either the true model or a model larger than the true model is selected. All plots are based on 500 simulation iterations as the selection procedure always selected a model with both truly non-linear effects and (potentially) further noise variables.
Appendix B: Further Application Results

![Graphs showing partial effects of estimated non-linear deviations for covariates lbph and pgg45.](image)

Figure 6: Partial effects of estimated non-linear deviations for the covariates lbph and pgg45.

Appendix C: Simulation Code

The R-code and link to the software to reproduce simulation application results can be found at [https://github.com/davidruegamer/inference_boosting](https://github.com/davidruegamer/inference_boosting).

Appendix D: Computation Time

In the following an estimation of computation time of our software for different model setups is given. As in Section ??, we use the same data generating process, assuming 4 signal variables and a SNR of 1. Note that we did not use parallelization when sampling from the space $\mathcal{R}_y$ and run-times can be roughly divided by the number of cores, $\varrho$ when using parallelization of $\varrho$ cores. We use $p_0 \in \{5, 50, 104\}$ noise variables and a grid from 1 to $\min(p_0 \cdot 10^2, 10^4)$ iterations, in which the optimal stopping iterations is searched for via CV.
Figure 7: Average computation time in hours for 5 simulation iterations of our selective inference approach for one test vector and different number of noise variables (x-axis) as well as number of observations (colour).