High-Dimensional Additive Modeling

Lukas Meier* Sara van de Geer Peter Bühlmann
Seminar für Statistik, ETH Zürich

June 25, 2008

Abstract

We propose a new sparsity-smoothness penalty for high-dimensional generalized additive models. The combination of sparsity and smoothness is crucial for mathematical theory as well as performance for finite-sample data. We present a computationally efficient algorithm, with provable numerical convergence properties, for optimizing the penalized likelihood. Furthermore, we provide oracle results which yield asymptotic optimality of our estimator for high-dimensional but sparse additive models. Finally, an adaptive version of our sparsity-smoothness penalized approach yields large additional performance gains.

1 Introduction

Substantial progress has been achieved over the last years in estimating high-dimensional linear or generalized linear models where the number of covariates p is much larger than sample size n. The theoretical properties of penalization approaches like the Lasso [22] are now well understood [11, 18, 27, 19, 1] and this knowledge has led to several extensions or alternative approaches like Adaptive Lasso [28], Relaxed Lasso [17], Sure Independence Screening [9] and graphical model based methods [4]. Moreover, with the fast growing amount of high-dimensional data in e.g. biology, imaging or astronomy, these methods have shown their success in a variety of practical problems. However, in many situations the conditional expectation of the response given the covariates may not be linear. While the most important effects may still be detected by a linear model, substantial improvements are sometimes possible by using a more flexible class of models. Recently, some progress has been made regarding high-dimensional additive model selection [2, 14, 21] and some theoretical results are available [21].

In this paper, we consider the problem of estimating a high-dimensional generalized additive model where $p \gg n$. An approach for high-dimensional additive modeling is described and analyzed in [21]. Our work is different in the following respects. (i) We use an approach which penalizes both the sparsity and the roughness. This is particularly important if a large number of basis functions is used for modeling the additive components. (ii) Our computational algorithm which builds upon the idea of a group-Lasso problem has rigorous convergence properties and thus, it is provably correct for finding the optimum of a penalized likelihood function. (iii) We provide oracle results which establish asymptotic optimality of the procedure.

*Corresponding autor. E-Mail: meier@stat.math.ethz.ch

2 Penalized Maximum Likelihood for Additive Models

We consider high-dimensional additive regression models with a continuous response $Y \in \mathbb{R}$ and p covariates $x^{(1)}, \ldots, x^{(p)} \in \mathbb{R}$ connected through the model

$$Y_i = c + \sum_{j=1}^{p} f_j\left(x_i^{(j)}\right) + \varepsilon_i,\ i = 1, \ldots, n,$$

where c is the intercept term, ε_i are i.i.d. random variables with mean zero and $f_j : \mathbb{R} \to \mathbb{R}$ are smooth univariate functions. For identification purposes we assume that all f_j are centered, i.e.

$$\sum_{i=1}^{n} f_j(x_i^{(j)}) = 0$$

for $j = 1, \ldots, p$. We consider the case of fixed design, i.e. we treat the predictors $x^{(1)}, \ldots, x^{(p)}$ as non-random.

With some slight abuse of notation we also denote by f_j the n-dimensional vector $(f_j(x_1^{(j)}), \ldots, f_j(x_n^{(j)}))^T$. For a vector $f \in \mathbb{R}^n$ we define $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} f_i^2$.

2.1 The Sparsity-Smoothness Penalty

In order to construct an estimator which encourages sparsity at the function level, penalizing the norms $\|f_j\|_n$ would be a suitable approach. Some theory for the case where a truncated orthogonal basis with $O(n^{1/5})$ basis functions for each component f_j is used has been developed in [21].

If we use a large number of basis functions, which is necessary to be able to capture some functions at high complexity, the resulting estimator will produce function estimates which are too wiggly if the underlying true functions are very smooth. Hence, we need some additional control or restrictions of the smoothness of the estimated functions. In order to get sparse and sufficiently smooth function estimates, we propose the sparsity-smoothness penalty

$$J(f_j) = \lambda_1 \sqrt{\|f_j\|_n^2 + \lambda_2 I^2(f_j)},$$

where

$$I^2(f_j) = \int \left(f_j''(x)\right)^2 dx$$

measures the smoothness of f_j. The two tuning parameters $\lambda_1, \lambda_2 \geq 0$ control the amount of penalization.

Our estimator is given by the following penalized least squares problem

$$\hat{f}_1, \ldots, \hat{f}_p = \underset{f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} \left\| Y - \sum_{j=1}^{p} f_j \right\|_n^2 + \sum_{j=1}^{p} J(f_j), \tag{1}$$

where $\mathcal{F}$ is a suitable class of functions and $Y = (Y_1, \ldots, Y_n)^T$ is the vector of responses. If we assume that Y is centered, we can omit an unpenalized intercept term and the nature of the objective function in (1) automatically forces the function estimates $\hat{f}_1, \ldots, \hat{f}_p$ to be centered.

Proposition 1. *Let $a, b \in \mathbb{R}$ such that $a < \min_{i,j}\{x_i^{(j)}\}$ and $b > \max_{i,j}\{x_i^{(j)}\}$. Let $\mathcal{F}$ be the space of functions that are twice continuously differentiable on $[a, b]$ and assume that there exist minimizers $\hat{f}_j \in \mathcal{F}$ of (1). Then the $\hat{f}_j$'s are natural cubic splines with knots at $x_i^{(j)}, i = 1, \ldots, n$.*

A proof is given in the Appendix. Hence, we can restrict ourselves to the finite dimensional space of natural cubic splines instead of considering the infinite dimensional space of twice continuously differentiable functions.

In the following subsection we illustrate the existence and the computation of the estimator.

2.2 Computational Algorithm

For each function f_j we use a cubic B-spline parameterization with a reasonable amount of knots or basis functions. A typical choice would be to use $K - 4 \asymp \sqrt{n}$ interior knots that are placed at the empirical quantiles of $x^{(j)}$. Hence, we parameterize

$$f_j(x) = \sum_{k=1}^{K} \beta_{j,k} b_{j,k}(x),$$

where $b_{j,k} : \mathbb{R} \to \mathbb{R}$ are the B-spline basis functions and $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,K})^T \in \mathbb{R}^K$ is the parameter vector corresponding to f_j. Based on the basis functions we can construct an $n \times pK$ design matrix $B = [B_1 \,|\, B_2 \,|\, \ldots \,|\, B_p]$, where B_j is the $n \times K$ design matrix of the B-spline basis of the jth predictor, i.e. $B_{j,il} = b_{j,l}(x_i^{(j)})$.

For twice continuously differentiable functions, the optimization problem (1) can now be re-formulated as

$$\hat{\beta} = \underset{\beta=(\beta_1,\ldots,\beta_p)}{\operatorname{argmin}} \|Y - B\beta\|_n^2 + \lambda_1 \sum_{j=1}^{p} \sqrt{\frac{1}{n}\beta_j^T B_j^T B_j \beta_j + \lambda_2 \beta_j^T \Omega_j \beta_j}, \tag{2}$$

where the $K \times K$ matrix Ω_j contains the inner products of the second derivatives of the B-spline basis functions, i.e.

$$\Omega_{j,kl} = \int b''_{j,k}(x) b''_{j,l}(x)\, dx$$

for $k, l \in \{1, \ldots, K\}$.

Hence, (2) can be re-written as a general Group Lasso problem [26]

$$\hat{\beta} = \underset{\beta=(\beta_1,\ldots,\beta_p)}{\operatorname{argmin}} \|Y - B\beta\|_n^2 + \lambda_1 \sum_{j=1}^{p} \sqrt{\beta_j^T M_j \beta_j}, \tag{3}$$

where $M_j = \frac{1}{n} B_j^T B_j + \lambda_2 \Omega_j$. By decomposing (e.g. using the Cholesky decomposition) $M_j = R_j^T R_j$ for some quadratic $K \times K$ matrix R_j and by defining $\tilde{\beta}_j = R_j \beta_j$, $\tilde{B}_j = B_j R_j^{-1}$, (3) reduces to

$$\hat{\tilde{\beta}} = \underset{\tilde{\beta}=(\tilde{\beta}_1,\ldots,\tilde{\beta}_p)}{\operatorname{argmin}} \left\| Y - \tilde{B}\tilde{\beta} \right\|_n^2 + \lambda_1 \sum_{j=1}^{p} \|\tilde{\beta}_j\|, \tag{4}$$

where $\|\tilde{\beta}_j\| = \sqrt{K}\|\tilde{\beta}_j\|_K$ is the Euclidean norm in $\mathbb{R}^K$. This is an ordinary Group Lasso problem for any fixed λ_2 and hence the existence of a solution is guaranteed. For λ_1

large enough, some of the coefficient groups $\beta_j \in \mathbb{R}^K$ will be estimated to be exactly zero. Hence, the corresponding function estimate will be zero. Moreover, there exists a value $\lambda_{1,max} < \infty$ such that $\hat{\tilde{\beta}}_1 = \ldots = \hat{\tilde{\beta}}_p = 0$ for $\lambda_1 \geq \lambda_{1,max}$. This is especially useful to construct a grid of λ_1 candidate values for cross-validation (usually on the log-scale).

Regarding the uniqueness of the identified components, we can make use of existing results of the Lasso. Define by $S(\tilde{\beta}; \tilde{B}) = \|Y - \tilde{B}\tilde{\beta}\|_n^2$. Similar to [20], the gradient $\nabla_{\tilde{\beta}} S(\tilde{\beta}; \tilde{B})$ is constant across all solutions of (4). In summary, we have the following Proposition.

Proposition 2. *If $pK \leq n$ and if $\tilde{B}$ has full rank, a unique solution of (4) exists. If $pK > n$, there exists a convex set of solutions of (4). Moreover, if $\|\nabla_{\tilde{\beta}_j} S(\hat{\tilde{\beta}}; \tilde{B})\| < \lambda_1$ then $\hat{\tilde{\beta}}_j = 0$ and all other solutions $\hat{\tilde{\beta}}_{other}$ satisfy $\hat{\tilde{\beta}}_{other,j} = 0$.*

By re-writing the original problem (1) in the form of (4), we can make use of already existing algorithms [16, 13, 26] to compute the estimator. Coordinate-wise approaches as in [16, 26] are efficient and have rigorous convergence properties. Thus, we are able to compute the estimator exactly, even if p is very large.

An example of estimated functions, from simulated data according to Example 1 in Section 3, is shown in Figure 1. For illustrational purposes we have also plotted the estimator which involves no smoothness penalty ($\lambda_2 = 0$). The latter clearly shows that for this example, the function estimates are "too wiggly" compared to the true functions.

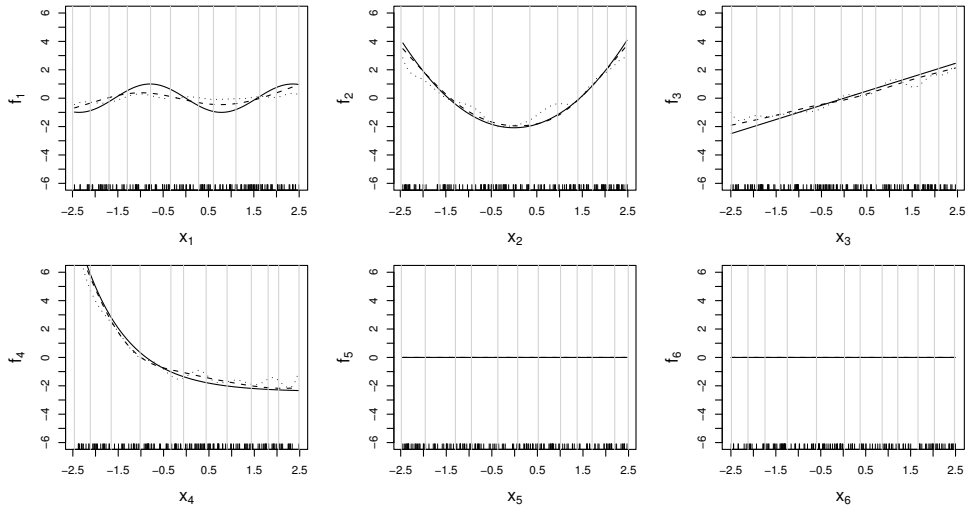

Figure 1: True functions f_j (solid) and estimated functions $\hat{f}_j$ (dashed) for the first 6 components of a simulation run of Example 1 in Section 3. Small vertical bars indicate original data and grey vertical lines knot positions. The dotted lines are the function estimates when no smoothness penalty is used, i.e. when setting $\lambda_2 = 0$.

Remark 1. *If we set $\lambda_2 = 0$, i.e. if we do not penalize the smoothness of the estimated functions, the block-matrices $\tilde{B}_j$ are orthogonal. This is exactly the parameterization that was proposed in [26]. Hence, this prediction type penalty can be interpreted as a Group Lasso penalty which is invariant under* all *transformations.*

Remark 2. *Two obvious alternative possibilities of our penalty would be to use either (i) $J(f_j) = \lambda_1\|f_j\|_n + \lambda_2 I(f_j)$ or (ii) $J(f_j) = \lambda_1\|f_j\|_n + \lambda_2 I^2(f_j)$. While proposal (i) also enjoys nice theoretical properties (see also Section 5.2), it is computationally more demanding, because it leads to a second order cone programming problem. Proposal (ii) basically leads again to a Group Lasso problem but appears to have theoretical drawbacks, i.e. the term $\lambda_2 I^2(f_j)$ is really needed within the square root.*

2.3 Oracle Results

We present now an oracle inequality for the penalized estimator. The proofs can be found in the Appendix.

We consider here a more general penalty of the form

$$J(f_j) = \lambda_1\sqrt{\|f_j\|_n^2 + \lambda_2 I^2(f_j)} + \lambda_3 I^2(f_j).$$

This penalty involves three smoothing parameters λ_1, λ_2 and λ_3. One may reduce this to a single smoothing parameter by choosing

$$\lambda_2 = \lambda_3 = \lambda_1^2,$$

(see Theorem 1 below). In the simulations however, the choice $\lambda_3 = 0$ turned out to provide slightly better results than the choice $\lambda_2 = \lambda_3$. With $\lambda_3 = 0$, the theory goes through provided the smoothness $I(\hat{f}_j)$ remains bounded in an appropriate sense.

We let f^0 denote the "true" regression function (which is not necessarily additive), i.e., we suppose the regression model

$$Y_i = f^0(x_i) + \varepsilon_i,$$

where $x_i = (x_i^{(1)}, \ldots, x_i^{(p)})^T$ for $i = 1, \ldots, n$, and where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random errors with $\mathbb{E}[\varepsilon_i] = 0$. Let f^* be a (sparse) additive approximation of f^0 of the form

$$f^*(x_i) = c^* + \sum_{j=1}^{p} f_j^*(x_i^{(j)}).$$

where we take $c^* = \mathbb{E}[\bar{Y}]$, $\bar{Y} = \sum_{i=1}^n Y_i/n$. The result of this subsection (Theorem 1) holds for any such f^* satisfying the compatibility condition below. Thus, one may invoke the optimal additive predictor among such f^*, which we will call the "oracle". For an additive function f, the squared distance $\|f - f^0\|_n^2$ can be decomposed into

$$\|f - f^0\|_n^2 = \|f - f_{add}^0\|_n^2 + \|f_{add}^0 - f^0\|_n^2,$$

where f_{add}^0 is the projection of f^0 on the space of additive functions. Thus, when f^0 is itself not additive, the oracle can be seen as the best sparse approximation of the projection f_{add}^0 of f^0.

The *active set* is defined as

$$\mathcal{A}_* = \{j : \ \|f_j^*\|_n \neq 0\}. \tag{5}$$

We will use a compatibility condition, in the spirit of the incoherence conditions used for proving oracle inequalities for the standard Lasso (see e.g. [1, 5, 6, 7, 23]). To avoid digressions, we will not attempt to formulate the most general condition. A discussion can be found in Section 5.1.

Compatibility condition *For a constant $0 < \phi_* \leq 1$, it holds that for all $\{f_j\}_{j=1}^p$,*

$$\sum_{j \in \mathcal{A}_*} \|f_j\|_n^2 \leq \|\sum_{j=1}^p f_j\|_n^2 / \phi_*^2.$$

Consider the general case where I is some semi-norm, e.g. as in Section 2.1. Write

$$f_j = g_j + h_j, \tag{6}$$

with g_j and h_j centered orthogonal functions, satisfying $I(h_j) = 0$ and $I(g_j) = I(f_j)$. The functions h_j are assumed to lie in a d-dimensional space. The entropy of $(\{g_j : I(g_j) = 1\}, \|\cdot\|_n)$ is denoted by $H_j(\cdot)$, see e.g. [25]. We assume that for all j,

$$H_j(\delta) \leq A\delta^{-2(1-\alpha)}, \ \delta > 0, \tag{7}$$

where $0 < \alpha < 1$ and $A > 0$ are constants. When $I^2(f_j) = \int \left(f_j''(x)\right)^2 dx$, the functions h_j are the linear part of f_j, i.e. $d = 1$. Moreover, one then has $\alpha = 3/4$ (see e.g. [25], Lemma 3.9).

Finally, we assume sub-Gaussian tails for the errors: for some constants L and M,

$$\max_i \mathbb{E}\left[\exp\left(\varepsilon_i^2/L\right)\right] \leq M. \tag{8}$$

The next lemma presents the behavior of the empirical process. We use the notation $(\varepsilon, f)_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)$ for the inner product. Define

$$\mathcal{S} = \mathcal{S}_1 \cap \mathcal{S}_2 \cap \mathcal{S}_3 \tag{9}$$

where

$$\mathcal{S}_1 = \left\{ \max_j \sup_{g_j} \left(\frac{2|(\varepsilon, g_j)_n|}{\|g_j\|_n^\alpha I^{1-\alpha}(g_j)} \right) \leq \xi_n \right\},$$

$$\mathcal{S}_2 = \left\{ \max_j \sup_{h_j} \left(\frac{2|(\varepsilon, h_j)_n|}{\|h_j\|_n} \right) \leq \xi_n \right\},$$

and

$$\mathcal{S}_3 = \{\bar{\varepsilon} \leq \xi_n\}, \ \bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i.$$

For an appropriate choice of ξ_n, the set $\mathcal{S}$ has large probability.

Lemma 1. *Assume (7) and (8). There exist constants c and C depending only on d, α, A, L, and M, such that for*

$$\xi_n \geq C\sqrt{\frac{\log p}{n}},$$

one has

$$\mathbb{P}(\mathcal{S}) \geq 1 - c\exp[-n\xi_n^2/c^2].$$

For $\alpha \in (0, 1)$, we define its "conjugate" $\gamma = 2(1-\alpha)/(2-\alpha)$. Recall that when $I^2(f_j) = \int \left(f_j''(x)\right)^2 dx$, one has $\alpha = 3/4$, and hence $\gamma = 2/5$.

We are now ready to state the oracle result for $\hat{f} = \hat{c} + \sum_{j=1}^p \hat{f}_j$ as defined in (1), with $\hat{c} = \bar{Y}$.

Theorem 1. *Take for $j = 1, \ldots, p$,*

$$J(f_j) = \lambda_1 \sqrt{\|f_j\|_n^2 + \lambda_2 I^2(f_j)} + \lambda_3 I^2(f_j),$$

with $\lambda_1 = \lambda^{\frac{2-\gamma}{2}}$ and $\lambda_2 = \lambda_3 = \lambda_1^2$, and with $2\sqrt{2}\xi_n \leq \lambda \leq 1$. Suppose the compatibility condition is met. Then on the set $\mathcal{S}$ given in (9), it holds that

$$\|\hat{f} - f_{add}^0\|_n^2 + \lambda^{\frac{2-\gamma}{2}} \sum_{j=1}^p \|\hat{f}_j - f_j^*\|_n$$

$$\leq 3\|f^* - f_{add}^0\|_n^2 + 4\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \left[I^2(f_j^*) + \frac{3}{\phi_*^2} \right] + \xi_n^2.$$

We remark that we did not attempt to optimize the constants given in Theorem 1, but rather looked for a simple explicit bound.

Remark 3. *In view of Lemma 1, one may take (under the conditions of this lemma) the smoothing parameter λ of order $\sqrt{\log p/n}$. When $I^2(f_j) = \int \left(f_j''(x)\right)^2 dx$, this gives $\lambda^{2-\gamma}$ of order $(\log p/n)^{4/5}$, which is up to the log-term the usual rate for estimating a twice differentiable function. If the oracle f^* has bounded smoothness $I(f_j^*)$ for all j, the rate is thus $p_{act}(\log p/n)^{4/5}$, with $p_{act} = |\mathcal{A}_*|$ being the number of active variables the oracle needs. This is, again up to the log-term, the same rate one would obtain if it was known beforehand which of the p functions are relevant.*

Remark 4. *The result implies that with large probability, the estimator selects a sup-set of the active functions, provided that the latter have enough signal (such kind of variable screening results have been established for the Lasso in linear and generalized linear models [24, 19]). More precisely, let $\mathcal{A}_0 = \{j : \|f_{add,j}^0\|_n \neq 0\}$ be the active set of f_{add}^0. Assume the compatibility condition holds for $\mathcal{A}_0$, with constant ϕ_0. Suppose also that for $j \in \mathcal{A}_0$, the smoothness is bounded, say $I(f_{add,j}^0) \leq 1$. Choosing $f^* = f_{add}^0$ in Theorem 1, tells us that on $\mathcal{S}$,*

$$\sum_{j=1}^p \|\hat{f}_j - f_{add,j}^0\|_n \leq 16\lambda^{\frac{2-\gamma}{2}} |\mathcal{A}_0|/\phi_0^2 + \xi_n^2.$$

Hence, if

$$\|f_{add,j}^0\|_n > 16\lambda^{\frac{2-\gamma}{2}} |\mathcal{A}_0|/\phi_0^2 + \xi_n^2, \ j \in \mathcal{A}_0,$$

we have (on $\mathcal{S}$), that the estimated active set $\{j : \|\hat{f}_j\|_n \neq 0\}$ contains $\mathcal{A}_0$.

3 Numerical examples

3.1 Simulations

In this section we investigate the empirical properties of the proposed estimator. We compare our approach with the Boosting approach of [2], where smoothing splines with low degrees of freedom are used as base learners; see also [3]. For $p = 1$, boosting with splines is known to be able to adapt to the smoothness of the underlying true function [2]. Generally, boosting is a very powerful machine learning method and a wide variety of software implementations are available, e.g. the R add-on package `mboost`.

We use a training set of n samples to train the different methods. An independent validation set of size $\lfloor n/2 \rfloor$ is used to select the prediction optimal tuning parameters λ_1 and λ_2. For boosting, the number of boosting iterations is used as tuning parameter. The shrinkage factor ν and the degrees of freedom *df* of the boosting procedure are set to their default values $\nu = 0.1$ and $df = 4$; see also [3].

By SNR we denote the signal-to-noise ratio, which is defined as

$$\text{SNR} = \frac{\text{Var}(f(X))}{\text{Var}(\varepsilon)},$$

where $f = f^0 : \mathbb{R}^p \to \mathbb{R}$ is the true underlying function.

A total of 100 simulation runs are used for each of the following settings.

3.1.1 Models

We use the following simulation models.

Example 1 ($n = 150$, $p = 200$, $p_{act} = 4$, SNR ≈ 15)
This example is similar to Example 1 in [21] and [12]. The model is

$$Y_i = f_1(x_i^{(1)}) + f_2(x_i^{(2)}) + f_3(x_i^{(3)}) + f_4(x_i^{(4)}) + \varepsilon_i,\ \varepsilon_i \text{ i.i.d. } N(0,1),$$

with

$$f_1(x) = -\sin(2x),\ f_2(x) = x_2^2 - 25/12,\ f_3(x) = x,$$
$$f_4(x) = e^{-x} - 2/5 \cdot \sinh(5/2).$$

The covariates are simulated from independent Uniform$(-2.5,\ 2.5)$ distributions. The true and the estimated functions of a simulation run are illustrated in Figure 1.

Example 2 ($n = 100$, $p = 1000$, $p_{act} = 4$, SNR ≈ 6.7)
As above but high-dimensional and correlated. The covariates are simulated according to a multivariate normal distribution with covariance matrix $\Sigma_{ij} = 0.5^{|i-j|}; i, j = 1, \ldots, p$.

Example 3 ($n = 100$, $p = 80$, $p_{act} = 4$, SNR $\approx 9\,(t = 0)$, $\approx 7.9\,(t = 1)$)
This is similar to Example 1 in [14] but with more predictors. The model is

$$Y_i = 5f_1(x_i^{(1)}) + 3f_2(x_i^{(2)}) + 4f_3(x_i^{(3)}) + 6f_4(x_i^{(4)}) + \varepsilon_i,\ \varepsilon_i \text{ i.i.d. } N(0, 1.74),$$

with

$$f_1(x) = x,\ f_2(x) = (2x-1)^2,\ f_3(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)}$$

and

$$f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x).$$

The covariates $x = (x^{(1)}, \ldots, x^{(p)})^T$ are simulated according to

$$x^{(j)} = \frac{W^{(j)} + tU}{1+t},\ j = 1, \ldots, p,$$

where $W^{(1)}, \ldots, W^{(p)}$ and U are i.i.d. Uniform(0, 1). For $t = 0$ this is the independent uniform case. The case $t = 1$ results in a design with correlation 0.5 between all covariates.

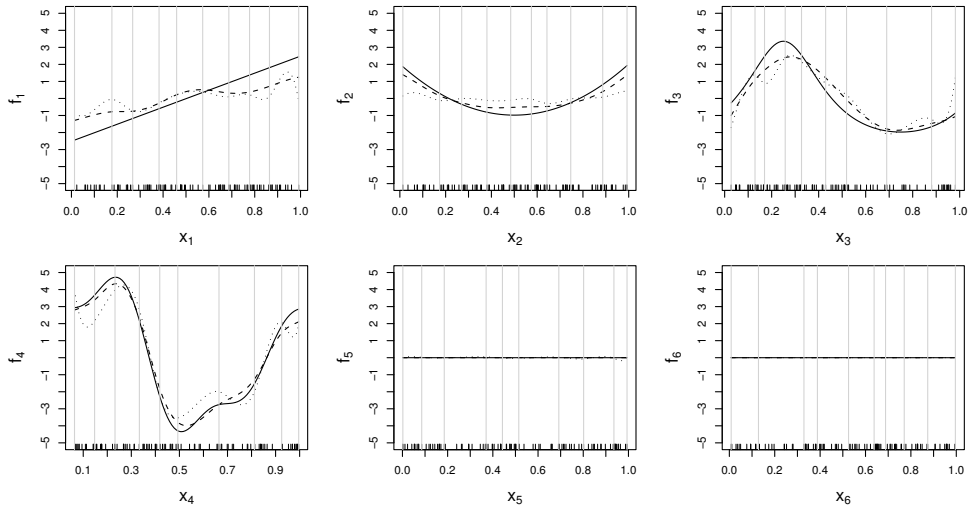

Figure 2: True functions f_j (solid) and estimated functions $\hat{f}_j$ (dashed) for the first 6 components of a simulation run of Example 3 ($t = 0$). Small vertical bars indicate original data and grey vertical lines knot positions. The dotted lines are the function estimates when no smoothness penalty is used, i.e. when setting $\lambda_2 = 0$.

The true functions and the first 6 estimated functions of a simulation run with $t = 0$ are illustrated in Figure 2.

Moreover, we also consider a "high-frequency" situation where we use $f_3(8x)$ and $f_4(4x)$ instead of $f_3(x)$ and $f_4(x)$. The corresponding signal-to-noise ratios for these models are SNR ≈ 9 for $t = 0$ and SNR ≈ 8.1 for $t = 1$.

Example 4 ($n = 100$, $p = 60$, $p_{act} = 12$, SNR $\approx 9\,(t = 0)$, $\approx 11.25\,(t = 1)$)
This is similar to Example 2 in [14] but with fewer observations. We use the same functions as in Example 3. The model is

$$\begin{aligned} Y_i &= f_1(x_i^{(1)}) + f_2(x_i^{(2)}) + f_3(x_i^{(3)}) + f_4(x_i^{(4)}) + \\ &\quad 1.5f_1(x_i^{(5)}) + 1.5f_2(x_i^{(6)}) + 1.5f_3(x_i^{(7)}) + 1.5f_4(x_i^{(8)}) + \\ &\quad 2f_1(x_i^{(9)}) + 2f_2(x_i^{(10)}) + 2f_3(x_i^{(11)}) + 2f_4(x_i^{(12)}) + \varepsilon_i, \end{aligned}$$

with ε_i i.i.d. $N(0, 0.5184)$. The covariates are simulated as in Example 3.

3.1.2 Performance Measures

In order to compare the prediction performances we use the mean squared prediction error

$$PE = \mathbb{E}_X[(\hat{f}(X) - f(X))^2]$$

as performance measure. The above expectation is approximated by a sample of 10,000 points from the distribution of X. In each simulation run we compute the ratio of the prediction performance of the two methods. Finally, we take the mean of the ratios over all simulation runs.

For variable selection properties we use the number of true positives (TP) and false positives (FP) at each simulation run. We report the average number over all runs to compare the different methods.

3.1.3 Results

The results are summarized in Table 1 and 2. The sparsity-smoothness penalty approach (SSP) has smaller prediction error than boosting, especially for the "high-frequency" situations. Because the weak learners of the boosting method only use 4 degrees of freedom, boosting tends to neglect or underestimate those components with higher oscillation. This can also be observed with respect to the number of true positives. By relaxing the smoothness penalty (i.e. choosing λ_2 small or setting $\lambda_2 = 0$), SSP is able to handle the high-frequency situations, at the cost of too wiggly function estimates for the remaining components. Using a different amount of regularization for sparsity and smoothness, SSP can work with a large amount of basis functions in order to be flexible enough to capture sophisticated functional relationships and, on the other side, to produce smooth estimates if the underlying functions are smooth.

With the exception of the high-frequency examples, the number of true positives (TP) is very similar for both methods. There is no clear trend with respect to the number of false positives (FP).

Model	PE_{SSP}/PE_{boost}
Example 1	0.93 (0.13)
Example 2	0.96 (0.10)
Example 3 ($t = 0$)	0.81 (0.13)
Example 3 ($t = 1$)	0.90 (0.19)
Example 3 "high-freq" ($t = 0$)	0.65 (0.11)
Example 3 "high-freq" ($t = 1$)	0.57 (0.10)
Example 4 ($t = 0$)	0.89 (0.10)
Example 4 ($t = 1$)	0.88 (0.13)

Table 1: Results of the different simulation models. Reported is the mean of the ratio of the prediction error of the two methods. SSP: Sparsity-Smoothness Penalty approach, boost: Boosting with smoothing splines. Standard deviations are given in parentheses.

Model	TP_{SSP}	FP_{SSP}	TP_{boost}	FP_{boost}
Example 1	4.00 (0.00)	24.30 (14.11)	4.00 (0.00)	22.18 (12.75)
Example 2	3.47 (0.61)	34.37 (17.38)	3.60 (0.64)	28.76 (20.15)
Example 3 ($t = 0$)	4.00 (0.00)	20.20 (9.30)	4.00 (0.00)	21.61 (10.90)
Example 3 ($t = 1$)	3.93 (0.29)	19.28 (9.61)	3.92 (0.27)	18.65 (8.35)
Example 3 "high-freq" ($t = 0$)	2.80 (0.78)	12.26 (7.61)	2.16 (0.94)	9.23 (9.74)
Example 3 "high-freq" ($t = 1$)	2.46 (0.85)	11.17 (8.50)	1.59 (1.27)	13.24 (13.89)
Example 4 ($t = 0$)	11.69 (0.56)	21.23 (6.85)	11.68 (0.57)	25.91 (9.43)
Example 4 ($t = 1$)	10.64 (1.15)	19.78 (7.51)	10.67 (1.25)	23.76 (9.89)

Table 2: Average values of the number of true (TP) and false (FP) positives. Standard deviations are given in parentheses.

3.2 Real Data

In this section we would like to compare the different estimators on real datasets.

3.2.1 Tecator

The `meatspec` dataset contains data from the Tecator Infratec Food and Feed Analyzer. It is for example available in the R add-on package `faraway` and on StatLib. The $p = 100$ predictors are channel spectrum measurements and are therefore highly correlated. A total of $n = 215$ observations are available.

The data is split into a training set of size 100 and a validation set of size 50. The remaining data are used as test set. On the training dataset, the first 30 principal components are calculated, scaled to unit variance and used as covariates in additive modeling. Moreover, the validation and the test dataset are transformed to correspond to the principal component of the training dataset. We fit an additive model to predict the logarithm of the fat content. This is repeated 50 times. For each split into training and test data we compute the ratio of the prediction errors from the SSP and boosting method on the test data, as in Section 3.1.2. The mean of the ratio over the 50 splits is 0.86, the corresponding standard deviation is 0.46. This indicates superiority of our sparsity-smoothness penalty approach.

3.2.2 Motif Regression

In motif regression problems [8], the aim is to predict gene expression levels or binding intensities based on information on the DNA sequence. For our specific dataset, from the Ricci lab at ETH Zurich, we have binding intensities Y_i of a certain transcription factor (TF) at 287 regions on the DNA. Moreover, for each region i, motif scores $x_i^{(1)}, \ldots, x_i^{(p)}, p = 196$ are available. A motif is a candidate for the binding site of the TF on the DNA, typically a 5–15bp long DNA sequence. The score $x_i^{(j)}$ measures how well the jth motif is represented in the ith region. The candidate list of motifs and their corresponding scores were created with a variant of the MDScan algorithm [15]. The main goal is here to find the relevant covariates.

We used 5 fold cross-validation to determine the prediction optimal tuning parameters, yielding 28 active functions. To assess the stability of the estimated model, we performed a nonparametric bootstrap analysis. At each of the 100 bootstrap samples, we fit the model with the fixed optimal tuning parameters from above. The two functions which appear most often in the bootstrapped model estimates are depicted in Figure 3. While the left-hand side plot shows an approximate linear relationship, the effect of the other motif seems to diminish for larger values. Indeed, `Motif.P1.6.26` is the true (known) binding site. A follow-up experiment showed that the TF does not directly bind to `Motif.P1.6.23`. Hence, this motif is a candidate for a binding site of a co-factor (another TF) and needs further experimental validation.

4 Extensions

4.1 Generalized Additive Models

Conceptually, we can also apply the sparsity-smoothness penalty from Section 2 to generalized linear models (GLM) by replacing the residual sum of squares $\|Y - \sum_{j=1}^{p} f_j\|_n^2$ by the corresponding negative log-likelihood function. We illustrate the method for logistic regression where $Y \in \{0, 1\}$. The negative log-likelihood as a function of the linear

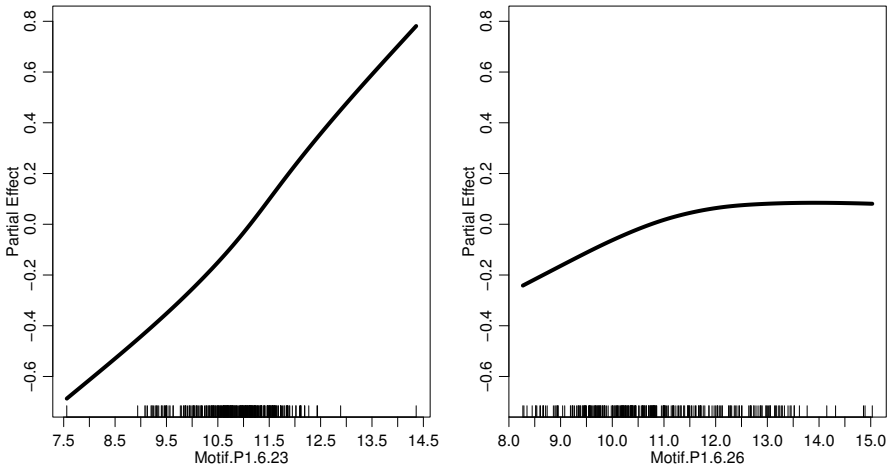

Figure 3: Estimated functions $\hat{f}_j$ of the two most stable motifs. Small vertical bar indicate original data.

predictor η and the response vector Y is

$$\ell(\eta, Y) = -\frac{1}{n}\sum_{i=1}^{n}[Y_i\eta_i - \log\{1+\exp(\eta_i)\}],$$

where $\eta_i = c + \sum_{j=1}^{p} f_j(x_i^{(j)})$. The estimator is defined as

$$\hat{c}, \hat{f}_1, \ldots, \hat{f}_p = \underset{c\in\mathbb{R}, f_1,\ldots,f_p\in\mathcal{F}}{\operatorname{argmin}} \ell\left(c + \sum_{j=1}^{p} f_j, Y\right) + \sum_{j=1}^{p} J(f_j). \tag{10}$$

This has a similar form as (1) with the exception that we have to explicitly include a (non-penalized) intercept term c. Using the same arguments as in Section 2 leads to the fact that for twice continuously differentiable functions, the solution can be represented as a natural cubic spline and that (10) leads again to a Group Lasso problem. This can for example be minimized with the algorithm of [16]. We illustrate the performance of the estimator in a small simulation study.

4.1.1 Small Simulation Study

Denote by $f : \mathbb{R}^p \to \mathbb{R}$ the true function of Example 2 in Section 3. We simulate the the linear predictor η as

$$\eta(X) = 1.5 \cdot (2 + f(X)),$$

where $X \in \mathbb{R}^p$ has the same distribution as in Example 2. The binary response Y is then generated according to a Bernoulli distribution with probability $1/(1+\exp(-\eta(X))$, which results in a Bayes risk of approximately 0.17. The sample size n is set to 100. The results for various model sizes p are reported in Table 3 and Table 4. The performance of the two methods is quite similar. SSP has a slightly lower prediction error. Regarding model selection properties, SSP has fewer false positives at the cost of slightly fewer true positives.

p	PE_{SSP}/PE_{boost}
250	0.93 (0.06)
500	0.96 (0.07)
1000	0.98 (0.05)

Table 3: Results of different model sizes p. Reported is the mean of the ratio of the prediction error of the two methods. SSP: Sparsity-Smoothness Penalty approach, boost: Boosting with smoothing splines. Standard deviations are given in parentheses.

p	TP_{SSP}	FP_{SSP}	TP_{boost}	FP_{boost}
250	2.94 (0.71)	22.81 (10.56)	3.09 (0.78)	29.67 (14.91)
500	2.56 (0.82)	24.92 (12.47)	2.80 (0.82)	31.41 (17.28)
1000	2.36 (0.84)	26.45 (14.88)	2.61 (0.71)	33.69 (19.54)

Table 4: Average values of the number of true (TP) and false (FP) positives. Standard deviations are given in parentheses.

4.2 Adaptivity

Similar to the Adaptive Lasso [28], we can also use different penalties for the different components, i.e. use a penalty of the form

$$J(f_j) = \lambda_1 \sqrt{w_{1,j}\|f_j\|_n^2 + \lambda_2 w_{2,j} I^2(f_j)},$$

where the weights $w_{1,j}$ and $w_{2,j}$ are ideally chosen in a data-adaptive way. If an initial estimator $\hat{f}_{j,init}$ is available, a choice would be to use

$$w_{1,j} = \frac{1}{\|\hat{f}_{j,init}\|_n^\gamma},\ w_{2,j} = \frac{1}{I(\hat{f}_{j,init})^\gamma}.$$

for some $\gamma > 0$. The estimator can then be computed similarly as described in Section 2.2. This allows for different degrees of smoothness for different components.

We have applied the adaptive estimator to the simulation models of Section 3. In each simulation run we use weights (with $\gamma = 1$) based on the ordinary sparsity-smoothness estimator. For comparison, we compute the ratio of the prediction error of the adaptive and the ordinary sparsity-smoothness estimator at each simulation run. The results are summarized in Table 5. Both the prediction error and the number of false positives can be decreased by a good margin in all examples. The number of true positives gets slightly decreased in some examples.

5 Mathematical Theory

5.1 On the compatibility condition

We will show that if the variables in the active set $\mathcal{A}_*$ in (5) are not too mutually dependent, and if the canonical dependence between the active and the non-active set is not perfect, then the compatibility condition is met.

Well-conditioned active set condition *We say that the active set $\mathcal{A}_*$ is well conditioned if for some constant $0 < \psi_* \leq 1$, and for all $\{f_j\}_{j\in\mathcal{A}_*}$,*

$$\sum_{j\in\mathcal{A}_*} \|f_j\|_n^2 \leq \|\sum_{j\in\mathcal{A}_*} f_j\|_n^2/\psi_*^2.$$

Model	$PE_{SSP;adapt}/PE_{SSP}$	TP	FP
Example 1	0.47 (0.13)	4.00 (0.00)	4.09 (4.63)
Example 2	0.63 (0.10)	3.31 (0.71)	6.12 (5.14)
Example 3 ($t = 0$)	0.53 (0.13)	4.00 (0.00)	4.64 (4.52)
Example 3 ($t = 1$)	0.63 (0.19)	3.81 (0.46)	5.04 (4.82)
Example 3 "high-freq" ($t = 0$)	0.87 (0.11)	2.28 (0.78)	2.98 (2.76)
Example 3 "high-freq" ($t = 1$)	0.91 (0.10)	1.69 (0.73)	2.59 (3.30)
Example 4 ($t = 0$)	0.77 (0.10)	11.21 (0.84)	8.18 (5.04)
Example 4 ($t = 1$)	0.88 (0.13)	9.73 (1.29)	7.93 (5.35)

Table 5: Results of the different simulation models. Reported is the mean of the ratio of the prediction error of the two methods and the average values of the number of true (TP) and false (FP) positives. SSP;adapt: Adaptive Sparsity-Smoothness Penalty approach, SSP: Ordinary Sparsity-Smoothness Penalty approach. Standard deviations are given in parentheses.

Writing f_j as a linear function of basis functions with coefficients β_j, e.g., as in Section 2.2,

$$f_j = B_j \beta_j,$$

with B_j the B-spline matrix of the jth predictor, one sees that ψ_*^2 can be taken as the smallest eigenvalue of the matrix

$$\left((B_j^T B_j)^{-1/2} (B_j^T B_k) (B_k^T B_k)^{-1/2} \right)_{j,k \in \mathcal{A}_*}.$$

The inner product between functions f and $\tilde{f}$ is denoted by $(f, \tilde{f})_n = \sum_{i=1}^n f(x_i)\tilde{f}(x_i)/n$. No perfect canonical dependence in our setup amounts to the following condition.

No perfect canonical dependence condition *We say that the active and non-active variables have no perfect canonical dependence, if for a constant $0 \leq \rho_* < 1$, and all $\{f_j\}_{j=1}^p$, we have for $f_{\text{in}} = \sum_{j \in \mathcal{A}_*} f_j$ and $f_{\text{out}} = \sum_{j \notin \mathcal{A}_*} f_j$, that*

$$\frac{|(f_{\text{in}}, f_{\text{out}})_n|}{\|f_{\text{in}}\|_n \|f_{\text{out}}\|_n} \leq \rho_*.$$

Again, writing $f_j = B_j \beta_j$, one sees that ρ_* can be taken as the canonical correlation between the linear space spanned by $\{B_j\}_{j \in \mathcal{A}_*}$ and the linear space spanned by $\{B_j\}_{j \notin \mathcal{A}_*}$. Note that the condition $\rho_* < 1$ allows for perfect linear dependencies between non-active B_j.

The next Lemma makes the link between the compatibility condition and the above two conditions.

Lemma 2. *Let $f = f_{\text{in}} + f_{\text{out}}$ satisfy*

$$\frac{|(f_{\text{in}}, f_{\text{out}})_n|}{\|f_{\text{in}}\|_n \|f_{\text{out}}\|_n} \leq \rho_* < 1.$$

Then

$$\|f_{\text{in}}\|_n^2 \leq \|f\|_n^2 / (1 - \rho_*^2).$$

Proof. Clearly,

$$\|f_{\text{in}}\|_n^2 \leq \|f\|_n^2 + 2|(f_{\text{in}}, f_{\text{out}})_n| - \|f_{\text{out}}\|_n^2.$$

Hence,

$$\|f_{\text{in}}\|_n^2 \leq \|f\|_n^2 + 2\rho_* \|f_{\text{in}}\|_n \|f_{\text{out}}\|_n - \|f_{\text{out}}\|_n^2 \leq \|f\|_n^2 + \rho_*^2 \|f_{\text{in}}\|_n^2.$$

□

Corollary 1. *A well-conditioned active set in combination with no perfect canonical dependence implies the compatibility condition from Section 2.3 with* $\phi_*^2 = \psi_*^2(1 - \rho_*^2)$.

Remark 5. *Let us define for all* $(j,k) \in \{1, \ldots, p\}$*, the canonical correlation*

$$\rho_{j,k} = \sup_{f_j, f_k} \frac{|(f_j, f_k)_n|}{\|f_j\|_n \|f_k\|_n}.$$

Let $R = (\rho_{j,k})_{j,k \in \mathcal{A}_*}$ *be the matrix of canonical correlations within the active set* $\mathcal{A}_*$*. Then* ψ_*^2 *can be taken as the smallest eigenvalue of* $2I_{id} - R$*, where* I_{id} *is the* $p_{act} \times p_{act}$ *identity matrix and* $p_{act} = |\mathcal{A}_*|$.

Remark 6. *Canonical dependence is about the dependence structure of variables. To compare, let* X_{in} *and* X_{out} *be two random variables, with joint density* p*, and with marginal densities* p_{in} *and* p_{out}*. Define for real-valued measurable functions* f_{in} *and* f_{out}*, of* X_{in} *and* X_{out} *respectively, the squared norms* $\|f_{in}\|^2 = \int f_{in}^2 p_{in}$*, and* $\|f_{out}\|^2 = \int f_{out}^2 p_{out}$*, and the inner product* $(f_{in}, f_{out}) = \int f_{in} f_{out} p$*. Assume the functions are centered:* $\int f_{in} p_{in} = \int f_{out} p_{out} = 0$*. Suppose that for some constant* ρ_**,*

$$\int \frac{p^2}{p_{in} p_{out}} \leq 1 + \rho_*^2.$$

Then one can easily verify that $|(f_{in}, f_{out})| \leq \rho_* \|f_{in}\| \|f_{out}\|$*. In other words, the no perfect canonical dependence condition is in this context the assumption that the density and the product density are, in* χ^2*-sense, not too far off.*

Remark 7. *It is clear that some condition on the dependence structure is needed. If two variables are highly correlated, our additive Lasso (with sparsity-smoothness penalty) should rather not include them both with opposite signs, i.e., the penalty hopefully prevents this. One may relax the canonical dependence condition in this spirit. First, one shows that only a subset of all possible* $\{f_j\}_{j=1}^p$ *needs to be considered. The relaxed condition is then that for this subset the correlation* $(f_{in}, f_{out})_n / (\|f_{in}\|_n \|f_{\text{out}})\|_n)$ *stays away from* -1*, i.e., that opposition does not pay off. We omit the details here to avoid digressions.*

5.2 On the choice of the penalty

In this paper, we have chosen the penalty in such a way that it leads to good theoretical behavior (namely the oracle inequality of Theorem 1), as well as to computationally fast, and in fact already existing, algorithms. The penalty can be improved theoretically, at the cost of computational efficiency and simplicity.

Indeed, a main ingredient from the theoretical point of view is that the randomness of the problem (the behavior of the empirical process) should be taken care of. Let us recall Lemma 1 which says that the set $\mathcal{S}$ has large probability, and on $\mathcal{S}$ all functions g_j satisfy

$$(\varepsilon, g_j)_n \leq \xi_n \|g_j\|_n^{\alpha} I^{1-\alpha}(g_j).$$

Our penalty was based on the inequality (which holds for any a and b positive)

$$a^\alpha b^{1-\alpha} \le \sqrt{a^2+b^2}.$$

More generally, it holds for any $q \ge 1$ that

$$a^\alpha b^{1-\alpha} \le (a^q + b^q)^{1/q}.$$

In particular, the choice $q = 1$ would be a natural one, and would lead to an oracle inequality involving $I(f_j^*)$ instead of the square $I^2(f_j^*)$ on the right-hand side in Theorem 1. The penalty $\lambda^{\frac{2-\gamma}{2}} \sum_{j=1}^p \|f_j\|_n + \lambda^{2-\gamma} \sum_{j=1}^p I(f_j)$, corresponding to $q = 1$, still involves convex optimization but which is much more involved and hence less efficient to be solved; see also Remark 2 in Section 2.2.

One may also use the inequality

$$a^\alpha b^{1-\alpha} \le a^2 + b^\gamma.$$

This leads to a "theoretically ideal" penalty of the from $\lambda^{2-\gamma} \sum_{j=1}^p I^\gamma(f_j) + \lambda \sum_{j=1}^p \|h_j\|_n$, where h_j is from (6). It allows to adapt to small values of $I(f_j^*)$. But clearly, as this penalty is non-convex, it may be computationally cumbersome. On the other hand, iterative approximations might prove to work well.

6 Conclusions

We present an estimator and algorithm for fitting sparse, high-dimensional generalized additive models. The estimator is based on a penalized likelihood. The penalty is new, as it allows for different regularization of the sparsity and the smoothness of the additive functions. It is exactly this combination which allows to derive oracle results for high-dimensional additive models. We also argue empirically that the inclusion of a smoothness-part into the penalty function yields much better results than having the sparsity-term only. Furthermore, we show that the optimization of the penalized likelihood can be written as a Group Lasso problem and hence, efficient coordinate-wise algorithms can be used which have provable numerical convergence properties.

We illustrate some empirical results for simulated and and real data. Our new approach with the sparsity and smoothness penalty is never worse and sometimes substantially better than L_2Boosting for generalized additive model fitting [2, 3]. Furthermore, with an adaptive sparsity-smoothness penalty method, large additional performance gains are achieved. With the real data about motif regression for finding DNA-sequence motifs, one among two selected "stable" variables is known to be true, i.e. it corresponds to a known binding site of a transcription factor.

A APPENDIX: Proofs

A.1 Proof of Proposition 1

Proof. Let $\hat{f}_1, \ldots, \hat{f}_p$ be a solution of (1) and assume that some or all $\hat{f}_j$ are not natural cubic splines with knots at $x_i^{(j)}$, $i = 1, \ldots, n$. By Theorem 2.2 in [10] we can construct natural cubic splines $\hat{g}_j$ with knots at $x_i^{(j)}$, $i = 1, \ldots, n$ such that $\hat{g}_j(x_i^{(j)}) = \hat{f}_j(x_i^{(j)})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Hence $\|Y - \sum_{j=1}^p \hat{g}_j\|_n^2 = \|Y - \sum_{j=1}^p \hat{f}_j\|_n^2$ and $\|\hat{g}_j\|_n^2 = \|\hat{f}_j\|_n^2$. But by Theorem 2.3 in [10], $I^2(\hat{g}_j) \le I^2(\hat{f}_j)$. Hence the minimizer of (1) can be represented by a natural cubic spline. □

A.2 Proof of Lemma 1

The result easily follows from Lemma 8.4 in [25], which we cite here for completeness.

Lemma 3. *Let $\mathcal{G}$ be a collection of functions $g : \{x_1, \ldots, x_n\} \to \mathbb{R}$, endowed with a metric induced by the norm $\|g\|_n = (\frac{1}{n}\sum_{i=1}^n g^2(x_i))^{1/2}$. Let $H(\cdot)$ be the entropy of $\mathcal{G}$. Suppose that*

$$H(\delta) \leq A\delta^{-2(1-\alpha)}, \ \forall\ \delta > 0.$$

Furthermore, let $\varepsilon_1, \ldots, \varepsilon_n$ be independent centered random variables, satisfying

$$\max_i \mathbb{E}\left[\exp\left(\varepsilon_i^2/L\right)\right] \leq M.$$

Then, for a constant c_0 depending on α, A, L and M, we have for all $T \geq c_0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{|2(\varepsilon, g)_n|}{\|g\|_n^\alpha} > \frac{T}{\sqrt{n}}\right) \leq c_0 \exp\left(-\frac{T^2}{c_0^2}\right).$$

Proof of Lemma 1. It is clear that $\{g_j/I(g_j)\} = \{g_j :\ I(g_j) = 1\}$. Hence, by rewriting and then using Lemma 3,

$$\sup_{g_j} \frac{|2(\varepsilon, g_j)_n|}{\|g_j\|_n^\alpha I^{1-\alpha}(g_j)} = \sup_{g_j} \frac{|2(\varepsilon, g_j/I(g_j))_n|}{\|g_j/I(g_j)\|_n^\alpha} \leq \frac{T}{\sqrt{n}},$$

with probability at least $1 - c_0\exp(-T^2/c_0^2)$. Thus, for $C_0^2 \geq 2c_0^2$ sufficiently large

$$\mathbb{P}\left(\max_j \sup_{g_j} \frac{|2(\varepsilon, g_j)_n|}{\|g_j\|_n^\alpha I^{1-\alpha}(g_j)} > C_0\sqrt{\frac{\log p}{n}}\right)$$

$$\leq pc_0 \exp\left(-\frac{C_0^2 \log p}{c_0^2}\right) \leq c_0 \exp\left(-\frac{C_0^2 \log p}{2c_0^2}\right).$$

In the same spirit, for some constant c_1 depending on L and M, it holds for all $T \geq c_1$, with probability at least $1 - c_1\exp(-T^2 d/c_1^2)$,

$$\sup_{h_j} \frac{|2(\varepsilon, h_j)_n|}{\|h_j\|_n} \leq T\sqrt{\frac{d}{n}},$$

where d is the dimension occurring in (6). This result is rather standard but also follows from the more general Corollary 8.3 in [25]. It yields that for $C_1^2 \geq 2c_1^2$, depending on d, L and M,

$$\max_j \sup_{h_j} \frac{|2(\varepsilon, h_j)_n|}{\|h_j\|_n} \leq C_1\sqrt{\frac{\log p}{n}},$$

with probability at least $1 - c_1\exp(-C_1^2 \log p/(2c_1^2))$.

Finally, it is obvious that for all C_2 and a constant c_2 depending on L and M,

$$\mathbb{P}(\bar{\varepsilon} > C_2\sqrt{\frac{\log p}{n}}) \leq 2\exp(-C_2^2 \log p/c_2^2).$$

Choosing $c_2 \geq 2$, the result now follows by taking $C = \max\{C_0, C_1, C_2\}$ and $c = c_0 + c_1 + c_2$. □

A.3 Proof of Theorem 1

We begin with three technical Lemmas.

Lemma 4. *For the decomposition in (6) it holds that*

$$\xi_n\|g_j - g_j^*\|_n^\alpha I^{1-\alpha}(g_j - g_j^*) + \xi_n\|h_j - h_j^*\|_n$$
$$\leq \frac{\lambda^{\frac{2-\gamma}{2}}}{2}\sqrt{\lambda^{2-\gamma}I^2(f_j - f_j^*) + \|f_j - f_j^*\|_n^2}$$

Proof. Note first that since $\xi_n \leq \lambda/(2\sqrt{2})$,

$$\xi_n\|g_j - g_j^*\|_n^\alpha I^{1-\alpha}(g_j - g_j^*) + \xi_n\|h_j - h_j^*\|_n$$
$$\leq \frac{\lambda}{2\sqrt{2}}\|g_j - g_j^*\|_n^\alpha I^{1-\alpha}(g_j - g_j^*) + \frac{\lambda}{2\sqrt{2}}\|h_j - h_j^*\|_n$$
$$\leq \frac{\lambda^{\frac{2-\gamma}{2}}}{2\sqrt{2}}\sqrt{\lambda^{2-\gamma}I^2(g_j - g_j^*) + \|g_j - g_j^*\|_n^2} + \frac{\lambda}{2\sqrt{2}}\|h_j - h_j^*\|_n$$
$$\leq \frac{\lambda^{\frac{2-\gamma}{2}}}{2\sqrt{2}}\sqrt{\lambda^{2-\gamma}I^2(g_j - g_j^*) + \|g_j - g_j^*\|_n^2} + \frac{\lambda^{\frac{2-\gamma}{2}}}{2\sqrt{2}}\|h_j - h_j^*\|_n,$$

since $\lambda \leq 1$.

We have

$$\sqrt{\lambda^{2-\gamma}I^2(g_j - g_j^*) + \|g_j - g_j^*\|_n^2} + \|h_j - h_j^*\|_n$$
$$\leq \sqrt{2\{\lambda^{2-\gamma}I^2(g_j - g_j^*) + \|g_j - g_j^*\|_n^2 + \|h_j - h_j^*\|_n^2\}}$$
$$= \sqrt{2}\sqrt{\lambda^{2-\gamma}I^2(g_j - g_j^*) + \|f_j - f_j^*\|_n^2}$$

where we used the orthogonality of $g_j - g_j^*$ and $h_j - h_j^*$. The result now follows from the equality $I(g_j - g_j^*) = I(f_j - f_j^*)$. □

Lemma 5. *We have*

$$\xi_n\|g_j\|_n^\alpha I^{1-\alpha}(g_j) + \xi_n\|h_j\|_n - J(f_j) \leq -J(f_j)/2.$$

Proof. By Lemma 4,

$$\xi_n\|g_j\|_n^\alpha I^{1-\alpha}(g_j) + \xi_n\|h_j\|_n \leq \frac{\lambda^{\frac{2-\gamma}{2}}}{2}\sqrt{\|f_j\|_n^2 + \lambda^{2-\gamma}I^2(f_j)}.$$

Hence,

$$\xi_n\|g_j\|_n^\alpha I^{1-\alpha}(g_j) + \xi_n\|h_j\|_n - J(f_j)$$
$$\leq -\lambda^{\frac{2-\gamma}{2}}\sqrt{\|f_j\|_n^2 + \lambda^{2-\gamma}I^2(f_j)} - \lambda^{2-\gamma}I^2(f_j) \leq -J(f_j)/2.$$

□

Lemma 6. *We have*

$$\xi_n\|g_j - g_j^*\|_n^\alpha I^{1-\alpha}(g_j - g_j^*) + \xi_n\|h_j - h_j^*\|_n + J(f_j^*) - J(f_j)$$
$$\leq \frac{3}{2}\lambda^{\frac{2-\gamma}{2}}\|f_j - f_j^*\|_n + 2\lambda^{2-\gamma}\Big[I^2(f_j^*) + 1\Big]. \tag{11}$$

Proof. We use the bound

$$J(f_j^*) - J(f_j) = \lambda^{\frac{2-\gamma}{2}} \sqrt{\|f_j^*\|_n^2 + \lambda^{2-\gamma} I^2(f_j^*)} + \lambda^{2-\gamma} I^2(f_j^*)$$

$$-\lambda^{\frac{2-\gamma}{2}} \sqrt{\|f_j\|_n^2 + \lambda^{2-\gamma} I^2(f_j)} - \lambda^{2-\gamma} I^2(f_j)$$

$$= \lambda^{\frac{2-\gamma}{2}} \sqrt{\|f_j^*\|_n^2 + \lambda^{2-\gamma} I^2(f_j^*)} - \lambda^{\frac{2-\gamma}{2}} \sqrt{\|f_j\|_n^2 + \lambda^{2-\gamma} I^2(f_j^*)}$$

$$+\lambda^{\frac{2-\gamma}{2}} \sqrt{\|f_j\|_n^2 + \lambda^{2-\gamma} I^2(f_j^*)} - \lambda^{\frac{2-\gamma}{2}} \sqrt{\|f_j\|_n^2 + \lambda^{2-\gamma} I^2(f_j)}$$

$$+\lambda^{2-\gamma} I^2(f_j^*) - \lambda^{2-\gamma} I^2(f_j)$$

$$\leq \lambda^{\frac{2-\gamma}{2}} \|f_j - f_j^*\|_n + \lambda^{2-\gamma} I(f_j - f_j^*) + \lambda^{2-\gamma} I^2(f_j^*) - \lambda^{2-\gamma} I^2(f_j)$$

$$\leq \lambda^{\frac{2-\gamma}{2}} \|f_j - f_j^*\|_n + \lambda^{2-\gamma} (I(f_j^*) + I^2(f_j^*)) + \lambda^{2-\gamma} I(f_j) - \lambda^{2-\gamma} I^2(f_j).$$

Using Lemma 4, it follows that

$$\xi_n \|g_j - g_j^*\|_n^\alpha I^{1-\alpha}(g_j - g_j^*) + \xi_n \|h_j - h_j^*\|_n + J(f_j^*) - J(f_j)$$

$$\leq \frac{\lambda^{2-\gamma}}{2} I(f_j - f_j^*) + \frac{\lambda^{\frac{2-\gamma}{2}}}{2} \|f_j - f_j^*\|_n$$

$$+\lambda^{\frac{2-\gamma}{2}} \|f_j - f_j^*\|_n + \lambda^{2-\gamma} (I(f_j^*) + I^2(f_j^*)) + \lambda^{2-\gamma} I(f_j) - \lambda^{2-\gamma} I^2(f_j)$$

$$\leq \frac{3}{2} \lambda^{\frac{2-\gamma}{2}} \|f_j - f_j^*\|_n + \lambda^{2-\gamma} \left(\frac{3}{2} I(f_j^*) + I^2(f_j^*) \right) + \lambda^{2-\gamma} \left(\frac{3}{2} I(f_j) - I^2(f_j) \right)$$

$$\leq \frac{3}{2} \lambda^{\frac{2-\gamma}{2}} \|f_j - f_j^*\|_n + 2\lambda^{2-\gamma} I^2(f_j^*) + 2\lambda^{2-\gamma},$$

where we used twice the (rough) inequality $(3/2)I \leq 1 + I^2$. □

Proof of Theorem 1. It holds that $\hat{c} = \bar{Y} (= \sum_{i=1}^n Y_i/n)$ and $c^* = \mathbb{E}\left[\bar{Y}\right]$. Thus, on $\mathcal{S}$, $|\hat{c} - c^*| \leq \xi_n$. Moreover,

$$\|\hat{f} - f^0\|_n^2 = |\hat{c} - c^*|^2 + \|(\hat{f} - \hat{c}) - (f^0 - c^*)\|_n^2.$$

To simplify the exposition (i.e., avoiding a change of notation), we may therefore assume $\hat{c} = c^*$ for the main part of the proof and add a ξ_n^2 to the final result. In the same spirit, we assume without loss of generality that $f^0 = f_{add}^0$.

On the set $\mathcal{S}$, it holds that

$$2|(\epsilon, \hat{f}_j - f_j^*)_n| \leq \xi_n \|\hat{g}_j - g_j^*\|_n^\alpha I^{1-\alpha}(\hat{g}_j - g_j^*)_n + \xi_n \|\hat{h}_j - h_j^*\|_n.$$

Thus, using the fact that the penalized loss at $\hat{f}$ is bounded by the penalized loss at f^*, we have

$$\|\hat{f} - f^0\|_n^2 + \sum_{j=1}^p J(\hat{f}_j) \leq 2|\sum_{j=1}^p (\epsilon, \hat{f}_j - f_j^*)_n| + \sum_{j=1}^p J(f_j^*) + \|f^* - f^0\|_n^2$$

$$\leq \xi_n \sum_{j=1}^p \|\hat{g}_j - g_j^*\|_n^\alpha I^{1-\alpha}(\hat{g}_j - g_j^*)_n + \xi_n \sum_{j=1}^p \|\hat{h}_j - h_j^*\|_n + \sum_{j=1}^p J(f_j^*) + \|f^* - f^0\|_n^2.$$

This implies

$$\|\hat{f} - f^0\|_n^2 - \|f^* - f^0\|_n^2 \leq i_{in} + ii_{out}, \tag{12}$$

where

$$i_{in} = \xi_n \sum_{j \in \mathcal{A}_*} \|\hat{g}_j - g_j^*\|_n^\alpha I^{1-\alpha}(\hat{g}_j - g_j^*)_n + \xi_n \sum_{j \in \mathcal{A}_*} \|\hat{h}_j - h_j^*\|_n$$

$$+ \sum_{j \in \mathcal{A}_*} J(f_j^*) - \sum_{j \in \mathcal{A}_*} J(\hat{f}_j),$$

and

$$ii_{out} = \xi_n \sum_{j \notin \mathcal{A}_*} \|\hat{g}_j\|_n^\alpha I^{1-\alpha}(\hat{g}_j) + \xi_n \sum_{j \notin \mathcal{A}_*} \|\hat{h}_j\|_n - \sum_{j \notin \mathcal{A}_*} J(\hat{f}_j).$$

In view of Lemma 6,

$$i_{in} \leq \frac{3}{2}\lambda^{\frac{2-\gamma}{2}} \sum_{j \in \mathcal{A}_*} \|f_j - f_j^*\|_n + 2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \Big[I^2(f_j^*) + 1 \Big].$$

Moreover, by Lemma 5,

$$ii_{out} \leq - \sum_{j \notin \mathcal{A}_*} J(\hat{f}_j)/2.$$

Hence,

$$i_{in} + ii_{out} + \frac{\lambda^{\frac{2-\gamma}{2}}}{2} \sum_{j=1}^{p} \|\hat{f}_j - f_j^*\|_n$$

$$\leq 2\lambda^{\frac{2-\gamma}{2}} \sum_{j \in \mathcal{A}_*} \|f_j - f_j^*\|_n + 2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \Big[I^2(f_j^*) + 1 \Big].$$

We now invoke the inequality

$$\sum_{j \in \mathcal{A}_*} \|\hat{f}_j - f_j^*\|_n \leq \sqrt{|\mathcal{A}_*|} \Big(\sum_{j \in \mathcal{A}_*} \|\hat{f}_j - f_j^*\|_n^2 \Big)^{1/2}$$

$$\leq \sqrt{|\mathcal{A}_*|} \|\hat{f} - f^*\|_n / \phi_*,$$

where the last inequality is simply the compatibility condition. This gives

$$i_{in} + ii_{out} + \frac{\lambda^{\frac{2-\gamma}{2}}}{2} \sum_{j=1}^{p} \|\hat{f}_j - f_j^*\|_n$$

$$\leq \frac{2}{\phi_*}\lambda^{\frac{2-\gamma}{2}} \sqrt{|\mathcal{A}_*|} \|\hat{f} - f^*\|_n + 2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \Big[I^2(f_j^*) + 1 \Big]$$

$$\leq \frac{2}{\phi_*}\lambda^{\frac{2-\gamma}{2}} \sqrt{|\mathcal{A}_*|} \|\hat{f} - f^0\|_n + \frac{2}{\phi_*}\lambda^{\frac{2-\gamma}{2}} \sqrt{|\mathcal{A}_*|} \|f^* - f^0\|_n + 2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \Big[I^2(f_j^*) + 1 \Big]$$

$$\leq \frac{1}{2}\|\hat{f} - f^0\|_n^2 + \frac{1}{2}\|f^* - f^0\|_n^2 + \frac{4}{\phi_*^2}\lambda^{2-\gamma}|\mathcal{A}_*| + 2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \Big[I^2(f_j^*) + 1 \Big]$$

$$= \frac{1}{2}\|\hat{f} - f^0\|_n^2 + \frac{1}{2}\|f^* - f^0\|_n^2 + 2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \Big[I^2(f_j^*) + 1 + \frac{2}{\phi_*^2} \Big]$$

Returning to (12), we see that

$$\|\hat{f} - f^0\|_n^2 - \|f^* - f^0\|_n^2 + \frac{\lambda^{\frac{2-\gamma}{2}}}{2} \sum_{j=1}^{p} \|\hat{f}_j - f_j^*\|_n$$

$$\leq \frac{1}{2}\|\hat{f} - f^0\|_n^2 + \frac{1}{2}\|f^* - f^0\|_n^2 + 2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \left[I^2(f_j^*) + 1 + \frac{2}{\phi_*^2} \right]$$

or

$$\frac{1}{2}\|\hat{f} - f^0\|_n^2 + \frac{\lambda^{\frac{2-\gamma}{2}}}{2} \sum_{j=1}^{p} \|\hat{f}_j - f_j^*\|_n \leq \frac{3}{2}\|f^* - f_0\|_n^2 + 2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \left[I^2(f_j^*) + 1 + \frac{2}{\phi_*^2} \right].$$

Because we assumed $\phi^* \leq 1$, we may simplify the last term to

$$2\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} \left[I^2(f_j^*) + \frac{3}{\phi_*^2} \right].$$

Finally, taking into account the rough bound ξ_n^2 for the estimation error for estimating c^*, the result follows. □

References

[1] P.J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2008. To appear.

[2] P. Bühlmann and B Yu. Boosting with the l2 loss: regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.

[3] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[4] Peter Bühlmann and Markus Kalisch. Variable selection for high-dimensional models: partial faithful distributions, strong associations and the pc-algorithm. Technical report, ETH Zürich, 2007.

[5] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation and sparsity via ℓ_1-penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory, COLT 2006. Lecture Notes in Artificial Intelligence 4005*, pages 379–391, Heidelberg, 2006. Springer Verlag.

[6] Florentina Bunea, Alexandre Tsybakov, and Marten H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[7] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

[8] Erin M. Conlon, X. Shirley Liu, Jason D. Lieb, and Jun S. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Science*, 100:3339 – 3344, 2003.

[9] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *Journal of Royal Statistical Society Series B*, 2008. To appear.

[10] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Number 58 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1994.

[11] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.

[12] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, 2004.

[13] Yuwon Kim, Jinseog Kim, and Yongdai Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2):375–390, 2006.

[14] Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297, 2006.

[15] X. Shirley Liu, Douglas L. Brutlag, and Jun S. Liu. An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, 20:835–839, 8 2002.

[16] Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, 70(1):53–71, 2008.

[17] Nicolai Meinshausen. Lasso with relaxation. *Computational Statistics and Data Analysis*, 52(1):374 – 393, 2007.

[18] Nicolai Meinshausen and Peter Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[19] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 2008. To appear.

[20] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[21] Pradeep Ravikumar, Han Liu, John Lafferty, and Larry Wasserman. Spam: Sparse additive models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1201–1208. MIT Press, Cambridge, MA, 2008.

[22] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

[23] S.A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36:614–645, 2008.

[24] Sara van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[25] Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, 2000.

[26] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67, 2006.

[27] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 2007. To appear.

[28] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.