Monotone Splines Lasso

Linn Cecilie Bergersen, Kukatharmini Tharmaratnam and Ingrid K. Glad

Department of Mathematics, University of Oslo

Abstract

We consider the problems of variable selection and estimation in nonparametric additive regression models for high-dimensional data. In recent years, several methods have been proposed to model nonlinear relationships when the number of covariates exceeds the number of observations by using spline basis functions and group penalties. Nonlinear monotone effects on the response play a central role in many situations, in particular in medicine and biology. We construct the monotone splines lasso (MS-lasso) to select variables and estimate effects using monotone spline (I-splines). The additive components in the model are represented by their I-spline basis function expansion and the component selection becomes that of selecting the groups of coefficients in the I-spline basis function expansion. We use a recent procedure, called cooperative lasso, to select sign-coherent groups, that is selecting the groups with either exclusively non-negative or non-positive coefficients. This leads to the selection of important covariates that have nonlinear monotone increasing or decreasing effect on the response. We also introduce an adaptive version of the MS-lasso which reduces both the bias and the number of false positive selections considerably. We compare the MS-lasso and the adaptive MS-lasso with other existing methods for variable selection in high dimensions by simulation and illustrate the method on two relevant genomic data sets. Results indicate that the (adaptive) MS-lasso has excellent properties compared to the other methods both by means of estimation and selection, and can be recommended for high-dimensional monotone regression.

Keywords: Cooperative lasso; High-dimensional data; I-splines; Lasso; Monotone regression; Nonparametric additive models

1 Introduction

In certain bio-medical applications it is important to assume that the relationship between an explanatory variable and the outcome is monotonically increasing or decreasing. Actually, every time linear regression is applied, an implicit monotonicity assumption is made. Monotone, but not necessarily linear relations appear typically for dose-response data, for example. It is also reasonable to assume that the relation between a disease and a risk factor is monotone, but not necessarily linear (Raftery and Richardson, 1996).

There has been a major effort in developing methods for monotone regression beyond the strictly linear regression models. These methods are usually concerned with classical situations in which the number of covariates $P$ does not exceed the number of observations $n$. In the last decade, the massive production of data sets in all areas of science and technology has turned high-dimensional regression problems, where $P$ is much larger than $n$, into one of the most active research areas. Very recently, an important contribution has appeared for monotone regression in high dimensions (Fang and Meinshausen, 2012). In this paper we develop another substantially different tool for this purpose.

When $P$ is larger than $n$, penalized regression methods handle the dimensionality problem by adding a penalty to the log-likelihood to be maximized. The lasso (Tibshirani, 1996) and its many variants (Zou (2006); van de Geer et al. (2011); Yuan and Lin (2006), Zou and
Hastie (2005); Meinshausen (2007) also have the advantage of setting some of the regression coefficients to zero, thus producing a sparse solution. These standard penalized regression methods limit the search to linear relationships between the covariates and the response. Recently nonparametric methods for high-dimensional regression have started to emerge. Recent papers (Meier et al. (2009), Huang et al. (2010); Ravikumar et al. (2009)) consider a generalized additive model (GAM) (Hastie and Tibshirani 1990) in combination with spline approximations. Given the observations \((y_i, x_i), i = 1, \ldots, n\), where \(y_i\) is the response and \(x_i = (x_{i1}, \ldots, x_{ip})\) is the vector of covariates for observation \(i\), the additive model is given by

\[
y_i = \beta_0 + \sum_{j=1}^{P} g_j(x_{ij}) + \epsilon_i.
\]

Here \(\beta_0\) is the intercept, the \(g_j\)'s are unknown functions to be estimated and \(\epsilon_i\) is the unobserved independent random error with mean zero and variance \(\sigma^2\). We assume \(E g_j^0(x_{ij}) = 0\), for \(1 \leq j \leq P\), where now \(x_{ij} = (x_{i1}, \ldots, x_{ij})\), to ensure unique identification of the \(g_j^0\)'s. In Meier et al. (2009), Bühlmann and van de Geer (2011), Huang et al. (2010) and Ravikumar et al. (2009) each nonparametric component \(g_j\) is represented by a linear combination of spline basis functions and the problem can be viewed in terms of a group lasso problem (Yuan and Lin 2006), hence selecting groups of spline basis functions representing relevant covariates. B-splines are popular for representing the covariates, because of their flexibility and minimal assumptions on the form of the function to be estimated. Combined with the group lasso, the framework becomes a highly flexible alternative to (standard) linear lasso-type methods. Our aim is to construct a new method that is nonparametric and flexible as above, but restricted to select and estimate monotone functions \(g_j\) only.

In simple regression problems, monotone increasing relationships are often modelled through isotonic regression (Barlow et al. 1972; Robertson et al. 1988) which is a nonparametric modelling technique restricting the regression function to be nondecreasing. Additive isotonic models assuming that each component effect in the additive model is isotonic, was presented in Bacchetti (1989). However, most literature on monotone and isotonic regression is limited to low dimensions. As mentioned, Fang and Meinshausen (2012) recently proposed Lasso Isotone (LISO) which combines estimation of nonparametric isotonic functions with ideas from sparse high-dimensional regression in an additive isotonic regression model, and is, to our knowledge, the only method feasible for monotone high-dimensional problems. In addition, using an adaptive liso approach, Fang and Meinshausen (2012) also present a way of fitting the model without assuming that all effects are either increasing or decreasing, that is, allowing for component effects of different sign. Isotonic regression is probably the best known method for preserving monotonicity, but has the disadvantage of producing step functions, which often has little biological plausibility, instead of smooth functions. For simple regression, it is possible to use an additional smoothing procedure in a second step to obtain a smooth function (He and Shi 1998; Tibshirani et al. 2011) also proposed nearly-isotonic regression which involves a penalty term controlling the level of monotonicity compared to the goodness of fit.

Another way of preserving monotonicity is to fit a smooth monotone function via monotone regression splines (Ramsay 1988; He and Shi 1998). While He and Shi (1998) proposed monotone B-spline smoothing based on a constrained least absolute deviation principle, Ramsay (1988) introduced integrated splines (I-splines), which essentially are integrated versions of M-splines that in combination with strictly positive coefficients will produce monotone increasing smooth functions. I-splines have previously been used in connection with a boosting technique to do monotonic regression in a multivariate model in Tutz and Leitenstorfer (2007). Also Meyer (2008) considers shape-restricted regression splines through I-splines, but only in the one-dimensional case.

In this paper we propose a new approach to fit nonparametric additive models under the assumption that each component effect \(g_j(x)\) is monotone. The \textit{monotone splines lasso} (MS-lasso) combines the idea of I-splines with the cooperative lasso (Chiquet et al. 2012), and is feasible in high-dimensional settings where the number of covariates \(P\) can exceed
the number of observations \( n \). The cooperative lasso is a lasso method where known groups of covariates are treated together, but differs from the standard group lasso \(^\text{[Yuan and Lin, 2006]}\) in that it assumes that the groups are sign-coherent. That is, the covariates inside a group are cooperating, so either the linear coefficients are all nonpositive, all nonnegative or all null inside a group. This can be combined with the idea of monotone I-splines by letting each covariate, represented via an I-splines basis, constitute a group in the cooperative lasso. Thus MS-lasso fits the additive nonparametric regression model with components that can be either nondecreasing, nonincreasing or of no effect. The important advantages of the MS-lasso are indeed that it is not restricted to only monotone increasing effects, that is, the monotone functions \( g^0_j \) can be either monotone increasing or decreasing in the same model, and also it is fitting smooth monotone functions to each \( g^0_j \). In this way we are more flexible than the linear model, but more restrictive than the pure nonlinear methods without any shape constraints. Our method is also biologically more relevant than the adaptive liso, in that we obtain smooth representations of the functions right away. We also suggest a two-step estimator, the adaptive MS-lasso, which leads to less bias and fewer false positives in the final model.

The paper is organized as follows. In Section 2 we present the MS-lasso and discuss some of its properties. Also the adaptive MS-lasso is presented, and we discuss connections to related methods. Section 3 is dedicated to simulation studies. In Section 4 we illustrate the use of our method in genomic data, before a final discussion is given in Section 5.

2 Monotone Splines Lasso

Suppose that each of the regression functions \( g^0_j \) in the additive model in (1) can be approximated by \( g_j \), a linear combination of \( m \) spline basis functions, that is,

\[
g_j(x) = \sum_{k=1}^{m} \beta_{jk} \phi_k(x), \quad 1 \leq j \leq P. \tag{2}
\]

Here \( \phi_k(\cdot) \) is the basis function and \( \beta_{jk} \) is the \( k \)-th spline coefficient for the \( j \)-th covariate. Note that for the standard linear regression model \( m = 1 \) and \( \phi_1(x) = x \).

The representation in (2) is used in Huang et al. (2010) with a cubic B-spline basis and six interior knots. Huang et al. (2010) combine this spline expansion with an (adaptive) group lasso penalty where the group structure is given by the basis functions for each covariate. This construction allows to perform selection of groups corresponding to important covariates.

Let \( \beta = (\beta_1, \ldots, \beta_P)^t \in \mathbb{R}^P_m \) with \( \beta_j = (\beta_{j1}, \ldots, \beta_{jm})^t \in \mathbb{R}^m \) for the \( j \)-th covariate when using \( m \) basis functions. The group lasso norm is

\[
||\beta||_{group} = \sum_{j=1}^{P} w_j ||\beta_j||
\]

where \( || \cdot || \) is the standard Euclidian norm and \( w_j > 0 \) are fixed weights for each covariate used to adapt the amount of penalty for each group. Typically \( w_j \) is the square root of the group size, to take care of situations with very different group sizes. When all covariates are represented by the same number of basis functions \( m, w_j = 1, \forall j \).

In the following construction of the MS-lasso we will introduce monotone splines in (2) and a special penalty, that guarantees monotonicity, in place of the group penalty above.

2.1 The MS-Lasso

For the MS-lasso, suppose without loss of generality that each \( x \) is transformed to take values in \([0, 1]\). For regression splines of order \( l \), we define knots \( 0 = t_1 = \ldots = t_i < \ldots < t_{K+l+1} = \ldots = t_{K+2l} = 1 \), where \( K \) is the fixed number of interior knots. Following Ramsay
we first define M-splines by $m = K + l$ M-splines basis functions. For order $l = 1$

$$M^{(1)}_k(x) = \begin{cases} \frac{1}{t_{k+1} - t_k}, & t_k \leq x \leq t_{k+1}, \\ 0 & \text{otherwise,} \end{cases}$$

for $k = 1, \ldots, K + 1$. For higher order $l$, the M-splines can be found by recursion from the lower order functions:

$$M^{(l)}_k(x) = \begin{cases} \frac{l((x-t_k)M^{(l-1)}_k(x)+(t_{k+1}-x)M^{(l-1)}_{k+1}(x))}{(l-1)(t_{k+1} - t_k)}, & t_k \leq x \leq t_{k+l}, \\ 0 & \text{otherwise.} \end{cases}$$

Finally Ramsay (1988) defines the integrated spline (I-spline) basis consisting of monotone functions by integrating

$$I^{(l)}_k(x) = \int_{t_k}^x M^{(l)}_k(u) du,$$

for $x \in [0, 1]$ and $k = 1, \ldots, K + l = m$. Based on this, for the MS-lasso we write (2) as

$$g_j(x) = \sum_{k=1}^m \beta_{jk} I^{(l)}_k(x), \quad 1 \leq j \leq P. \tag{3}$$

As each spline function $I^{(l)}_k(x)$ is monotone, monotonicity in $g_j(x)$ is achieved by restricting the spline coefficients to be of the same sign. An I-spline with all $\beta_{jk} \geq 0$, $\forall k$ will produce a monotone nondecreasing function $g_j(x)$, while an I-spline with $\beta_{jk} \leq 0$, $\forall k$ will produce a nonincreasing function. The cooperative lasso (Chiquet et al., 2012) is designed especially for problems of estimation and selection of parameters in high dimensions where the group structure is known and sign-coherence is assumed within the group.

We are centering the basis functions to satisfy $Eg_j(x_j) = 0$, for $1 \leq j \leq P$. Let $z_{ijk} = I_k(x_{ij}) - I_{jk}$ be the centered I-spline basis function, where $I_{jk} = \frac{1}{n} \sum_{i=1}^n I_k(x_{ij})$. Let $Z = (Z_1, \ldots, Z_P)$ be the $n \times (Pm)$ design matrix where all covariates are represented by a centered I-spline basis and $Z_j$ is the $n \times m$ matrix for the $j$th covariate. We use the centered response vector $\mathbf{y}$ of length $n$. Then the MS-lasso estimates of $\beta$ are defined by minimizing the objective function with respect to $\beta$, giving

$$\hat{\beta}^{MS} = \arg\min_{\beta \in \mathbb{R}^m} \left\{ \frac{1}{2} ||\mathbf{y} - Z\beta||^2 + \lambda||\beta||_{coop} \right\}, \tag{4}$$

where $\lambda \geq 0$ decides the amount of shrinkage and is common to all groups. Here

$$||\beta||_{coop} = ||\beta^+||_{group} + ||\beta^-||_{group} = \sum_{j=1}^P \sum_{k=1}^m w_j \left( ||\beta^+_{jk}|| + ||\beta^-_{jk}|| \right) \tag{5}$$

is the cooperative lasso norm with $\beta^+ = (\beta^+_1, \ldots, \beta^+_P)^t$ and $\beta^- = (\beta^-_1, \ldots, \beta^-_P)^t$, where $\beta^+_j = (\beta_{j1}, \ldots, \beta_{jm})^t$ and $\beta^-_j = (\beta_{j1}, \ldots, \beta_{jm})^t$. Hence $\beta^+$ and $\beta^-$ are the positive and negative parts of $\beta$, that is, $\beta^+_{jk} = \max(0, \beta_{jk})$ and $\beta^-_{jk} = \max(0, -\beta_{jk})$ respectively.

For simplicity we will use the I-splines basis of order $l = 2$ which has the closed form

$$I^{(2)}_k(x) = \begin{cases} 0, & x \leq t_k, \\ \frac{(x-t_k)^2}{(t_{k+1} - t_k)(t_{k+2} - t_k)}, & t_k < x \leq t_{k+1}, \\ 1 - \frac{(t_{k+2} - x)^2}{(t_{k+2} - t_k)(t_{k+2} - t_{k+1})}, & t_{k+1} < x \leq t_{k+2}, \\ 1, & x > t_{k+2}, \end{cases}$$

such that $g_j(x)$ is defined by letting $l = 2$ in (3). Finally the estimated monotone effects can be found as

$$\hat{g}_j(x) = \sum_{k=1}^m \beta_{jk}^{MS} I^{(2)}_k(x), \quad 1 \leq j \leq P.$$
The estimates \( \hat{\beta}^{MS} \) can be obtained by using the \( n \times (Pm) \) design matrix \( Z \) of centered I-spline basis function expansions in the cooperative lasso algorithm through the R-package scoop available at http://stat.genopole.cnrs.fr/logiciels/scoop.

2.2 Properties

In the following, we assume that each of the true functions \( g_0^j \) can be represented as a monotone I-spline function \( g_j \) as in (3) with \( m \) basis functions. To study the properties of the MS-lasso, we need some standard assumptions on the joint distribution of \( (Z, Y) \), which are required to guarantee the convergence of empirical covariances. We assume that the columns of \( Z \) are centered (hence we have zero mean random variables and \( E[ZZ^T] \) is the covariance matrix of \( Z \)), that \( Z \) and \( Y \) have finite fourth order moments, and that the covariance matrix of \( Z \) is invertible. Additionally, we need a specific assumption for the cooperative lasso, namely that in the sign-incoherent groups, all coefficients are nonzero. Furthermore, we need two suitable variations of the irrepresentable condition for the cooperative lasso. All of these assumptions are given in detail in (A1)-(A5) in Section 3 in Chiquet et al. (2012). Under these assumptions, the consistency of the cooperative lasso is assured for fixed \( P \). All required assumptions are also satisfied in our setting to achieve consistency for the MS-lasso for fixed \( P \) and fixed number of spline basis functions \( m \). We can re-write the optimization problem in (4) in the following normalized form,

\[
\hat{\beta}^{MS}_n = \arg\min_{\beta \in \mathbb{R}^{Pm}} \left\{ \frac{1}{2n} ||y - Z\beta||^2 + \lambda_n ||\beta||_{coop} \right\},
\]

(6)

where \( \lambda_n = \lambda/n \). Using the results in Theorem 2 in Chiquet et al. (2012) for fixed \( P \), the MS-lasso estimator \( \hat{\beta}^{MS}_n \) is asymptotically unbiased and has the property of exact support recovery,

\[
\hat{\beta}^{MS}_n \xrightarrow{P} \beta \quad \text{and} \quad \mathbb{P}(S(\hat{\beta}^{MS}_n) = S) \rightarrow 1,
\]

(7)

for every sequence \( \lambda_n \) such that \( \lambda_n = \lambda_0 n^{-\gamma}, \gamma \in (0, 1/2) \) and where \( S = \{j, \beta_j \neq 0\} \). It follows directly from (7) that \( \hat{g}_j \xrightarrow{P} g_j \). Under the assumption that \( g^0_j \) is in the space of monotone I-spline functions with \( m \) basis functions, we have that \( \hat{g}_j \xrightarrow{P} g^0_j \) and \( \mathbb{P}(S^*(\hat{g}_j) = S^*) \rightarrow 1 \), where \( S^* = \{j, g^0_j \neq 0\} \). That is, the difference between the estimated spline function \( \hat{g}_j \) and the true spline function \( g^0_j \) for each covariate in model (1) converges to zero in probability, and the nonzero components are selected correctly with probability converging to one.

2.3 The Adaptive MS-Lasso

Often, a second adaptive step is introduced for the lasso (Zou, 2006) and the group lasso (Bühlmann and van de Geer, 2011) procedures. This involves replacing the penalties by a re-weighted version based on an initial estimator of the regression coefficients. In the high-dimensional case, the lasso or group lasso themselves can be used to obtain the initial estimator. This idea can also be applied to obtain the adaptive MS-lasso by using the penalty in (5) and letting

\[
\hat{\beta}^{AMS} = \arg\min_{\beta \in \mathbb{R}^{Pm}} \left\{ \frac{1}{2} ||y - Z\beta||^2 + \lambda \sum_{j=1}^{P} w_j (||\hat{\beta}^+_j|| + ||\hat{\beta}^-_j||) \right\},
\]

(8)

where

\[
w_j = \begin{cases} 
||\hat{\beta}^{MS}_j||^{-1}, & \text{if } ||\hat{\beta}^{MS}_j|| > 0 \\
\infty, & \text{if } ||\hat{\beta}^{MS}_j|| = 0.
\end{cases}
\]

This assumes that the covariates are represented with the same number of basis functions.
Table 1: Overview of selection procedures.

| Method       | Penalty                                      | Basis                  | Selection            |
|--------------|----------------------------------------------|------------------------|----------------------|
| MS-lasso     | $\lambda \sum_{j=1}^{P} (||\beta_j^+|| + ||\beta_j^-||)$ | Monotone I-splines basis, $I^{(2)}_k(x)$ | Nonlinear monotone effects |
| AMS-lasso    | $\lambda \sum_{j=1}^{P} w_j (||\beta_j^+|| + ||\beta_j^-||)$ | I-splines basis, $I^{(2)}_k(x)$ | Nonlinear effects |
| Ad. liso     | $\lambda \sum_{j=1}^{P} \Delta(g_j)$        | Step function basis    | Monotone effects (not smooth) |
| Lasso        | $\lambda \sum_{j=1}^{P} |\beta_j|$                           | Original data, $X$    | Linear effects       |
| Ad. lasso    | $\lambda \sum_{j=1}^{P} w_j |\beta_j|$                        | B-splines basis       | Nonlinear effects    |
| BS-lasso     | $\lambda \sum_{j=1}^{P} w_j ||\beta_j||$    | B-splines basis $B^{(2)}_k(x)$ | Nonlinear effects |

The intuitive idea and motivation are the same as for the adaptive lasso and the adaptive group lasso. For example, the adaptive (group) lasso procedures have the property that the solution is at least as sparse as the initial estimator, and can therefore be used to reduce the number of false positives compared to the standard initial procedures. The adaptive procedures also penalize less for components with large initial estimators, implying less biased estimates than for the standard procedures.

2.4 Other Selection Procedures

We compare the performance of the MS-lasso and the adaptive MS-lasso with four other methods which are doing variable selection in high-dimensional regression. Although all of them are suitable in the high-dimensional setting, comparison is not completely fair as the scopes of the various methods are very different. That is, the underlying assumptions on the component functions are different and will lead to different selections depending on what method we use. It is, however, important to have a clear vision of what are their differences and when each of them should be applied depending on the problem.

Table 1 shows an overview of the different methods and the corresponding basis functions and penalties that are used. In Table 1, MS-lasso and AMS-lasso are the proposed methods, Lasso and Ad. lasso are standard and adaptive lasso respectively, while Ad. liso is the adaptive LISO in Fang and Meinshausen (2012), and BS-lasso is the adaptive group lasso using B-splines in Huang et al. (2010). It is the combination of type of penalty and representation of data that defines the assumptions we make on the component effects. For example, the lasso and the adaptive lasso together with the original data measurements, will be restricted to important linear effects, while the BS-lasso also allows for components describing the response nonlinearly without any restrictions on the shapes in the estimated nonzero effects. Fang and Meinshausen (2012) proposed the adaptive liso for additive isotonic regression models. The adaptive liso considers a two stage procedure, first calculating the initial fit using the liso algorithm. In the second stage, they conduct a liso procedure again with covariate weights based on the initial fit of the model. Finally, the MS-lasso and the adaptive MS-lasso select important monotone, but possibly nonlinear effects with the combination of monotone I-splines and the cooperative lasso penalty.
2.5 Choice of Tuning Parameter

The MS-lasso and the other methods listed in Table 1 all require selection of the penalty parameter $\lambda$ that controls the amount of shrinkage and hence the number of variables selected. To choose $\lambda$ we use K-fold cross-validation on a grid of $\lambda$ values defining the set of candidate models. K-fold cross-validation, typically with $K = 10$, involves splitting the data into $K$ folds. Leaving one fold out at a time, the candidate model is fitted on the remaining $9/10$ of the data while predicting for the left out fold. We choose the model minimizing an estimate of the prediction mean squared error,

$$CV(\lambda) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in f_k} (y_i - \hat{y}_{i,k}^{\lambda})^2,$$

where $f_k$ is the set of indices of the samples in fold $k$, and $\hat{y}_{i,k}^{\lambda}$ is the fitted predicted value for observation $i$ when fold $k$ involving observation $i$ is left out of the estimation. For the adaptive approaches, both the initial estimator and the final adaptive estimator can be estimated with $\lambda$ chosen by cross-validation.

3 Simulation Studies

To investigate the finite-sample performance of our method, we report the results from several simulation experiments. In all experiments we use the MS-lasso and the adaptive MS-lasso to estimate the component effects $g_j$ and compare with the methods listed in Table 1. For the monotone splines methods we use a monotone I-splines basis of order two and with six evenly distributed knots for all functions $g_j$. For the BS-lasso we use a quadratic B-spline basis, also with six evenly distributed knots. For all methods we use 10-fold cross-validation to find the optimal $\lambda$ for prediction.

We generate $w_{ij},\ldots,w_{iP},u_i$ and $v_i$ independently from $N(0,1)$ truncated to $[0,1]$. Then, following Huang et al. (2010), the covariates are generated as follows,

$$x_{ij} = \frac{w_{ij} + tu_i}{1 + t} \quad \text{for } j \in A, \quad x_{ij} = \frac{w_{ij} + tu_i}{1 + t} \quad \text{for } j \not\in A,$$

such that if $A$ is the set of components in the true model, the nonzero and zero components are independent, and the correlation among the covariates within the active and nonactive set respectively is controlled by $t$. In all situations the number of replications is 100, $P = 1000$, and $n = 50$.

To define the true models we consider

$$g_1^a(x) = -\exp(x^2), \quad g_2^a(x) = -\log(x + 0.1), \quad g_3^a(x) = 2\tanh(20x^2) + 0.5\exp(x^3),$$

$$g_4^a(x) = \frac{2\exp(10x - 5)}{1 + \exp(10x - 5)}, \quad g_5^a(x) = 2x.$$

Finally, the response variable is generated from the following models;

Model A $y_i = g_1^a(x_{i1}) + g_2^a(x_{i2}) + g_3^a(x_{i3}) + g_4^a(x_{i4}) + \epsilon_i$,

Model B $y_i = g_1^a(x_{i1}) + g_2^a(x_{i2}) + g_3^a(x_{i3}) + g_4^a(x_{i4}) + \epsilon_i$,

where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma^2$ is chosen to control the signal to noise ratio (SNR). Hence Model A contains only nonlinear monotone effects, while in Model B the fourth component is exchanged with a strictly linear effect.

Variable selection is compared in terms of the number of experiments for which the method selects each of the true components in the true model, as well as the overall number of true positives (TP) and false positives (FP). We also evaluate the fitted functions for the components in the true model by computing the mean squared error (MSE) between the true function value and the fitted value in the observed points.
To compare the selection performance, the proportion of correct selections in each component in the true model is reported together with the average number of true and false positives. For estimation, MSE between the fitted function and the true function, averaged over the 100 simulated data sets, is reported.

### Selection

| Method      | $g_1^0$  | $g_2^0$  | $g_3^0$  | $g_4^0$  | TP      | FP      |
|-------------|----------|----------|----------|----------|---------|---------|
| MS-lasso    | 1.00 (0.00) | 0.89 (0.31) | 1.00 (0.00) | 1.00 (0.00) | 3.89 (0.31) | 17.72 (9.84) |
| Ad. MS-lasso| 0.98 (0.14) | 0.87 (0.34) | 1.00 (0.00) | 1.00 (0.00) | 3.85 (0.41) | 2.97 (3.05) |
| Lasso       | 0.85 (0.36) | 0.72 (0.45) | 1.00 (0.00) | 1.00 (0.00) | 3.57 (0.62) | 25.01 (11.85) |
| Ad. lasso   | 0.81 (0.39) | 0.68 (0.47) | 1.00 (0.00) | 1.00 (0.00) | 3.49 (0.67) | 18.40 (7.72) |
| Ad. liso    | 0.39 (0.49) | 0.98 (0.14) | 1.00 (0.00) | 1.00 (0.00) | 3.37 (0.51) | 5.81 (2.46) |
| BS-lasso    | 0.00 (0.00) | 0.04 (0.20) | 0.23 (0.42) | 0.93 (0.26) | 1.20 (0.62) | 1.09 (1.87) |

### Estimation

| Method      | $g_1^0$  | $g_2^0$  | $g_3^0$  | $g_4^0$  |
|-------------|----------|----------|----------|----------|
| MS-lasso    | 0.06 (0.03) | 0.17 (0.08) | 0.15 (0.06) | 0.14 (0.06) |
| Ad. MS-lasso| 0.02 (0.03) | 0.07 (0.10) | 0.03 (0.02) | 0.03 (0.04) |
| Lasso       | 0.11 (0.04) | 0.26 (0.06) | 0.35 (0.07) | 0.21 (0.07) |
| Ad. lasso   | 0.09 (0.05) | 0.22 (0.09) | 0.28 (0.06) | 0.15 (0.07) |
| Ad. liso    | 0.12 (0.05) | 0.07 (0.06) | 0.08 (0.04) | 0.05 (0.02) |
| BS-lasso    | 0.16 (0.00) | 0.32 (0.03) | 0.62 (0.22) | 0.15 (0.13) |

### 3.1 Model A

For Model A we consider several situations including both independent ($t = 0$) and dependent covariates ($t = 1$), as well as two typical values for the signal to noise ratio; $SNR \approx 4$ and $SNR \approx 2$. Table 2 reports the results for all six methods in experiments with $SNR \approx 4$ and $t = 0$ in expression (9). For this situation we observe that the MS-lasso is able to select the four components that are really in the true model in nearly all simulation runs, hence the number of true positives is close to four. The MS-lasso has, however, a considerable number of false positives. Introducing an adaptive step reduces the number of false positives and the adaptive MS-lasso is best in recovering the true model. The two methods estimating linear effects, the standard lasso and the adaptive lasso, are also quite good in selecting the true nonzero components, but not as good as the MS-lasso methods and none of them are able to reduce the number of false positives sufficiently. The strongest competitor in this case, is the adaptive liso which performs well in variable selection. However, it is not better than the adaptive MS-lasso in terms of TP and FP. Finally the BS-lasso performs the worst; three of the true nonzero components are selected in less than 25% of the cases.

We also evaluate the estimation error which is given in the lower part of Table 2. Comparing the MS-lasso and adaptive MS-lasso with their linear competitors, the lasso and the adaptive lasso respectively, we see that the two methods allowing for a nonlinear monotone relationship, estimate the effect of the components with more accuracy. Comparing the MSE for each of the four components individually, we see that the adaptive MS-lasso does much better than all of the other methods. This can also be observed directly from Figure 1 and 2 which show the estimated component functions $\hat{g}_j, j = 1, \ldots, 4$ for all 100 replications. More specifically, the curves for the MS-lasso and the adaptive MS-lasso are plotted in the four upper panels of Figure 1, while the curves for the lasso and the adaptive lasso are plotted in the four lower panels. Figure 2 shows estimated curves for the BS-lasso and the adaptive liso. It is important to note that in Figure 1, the nonadaptive and adaptive version of each procedure are plotted in the same panels to illustrate the advantage of introducing an adaptive step. This might give the false impression that there is more variability in the
estimated curves, hence it is important to have in mind that in Figure 1 the estimated curves come from $2 \times 100$ simulation runs and should be considered separately as indicated by color. Obviously the adaptive MS-lasso captures the true shape of the functions better than the nonadaptive MS-lasso, but also better than all the other methods. Note that the adaptive liso also seems to capture the overall shape of the functions, but the estimation error becomes larger than for the MS-lasso because the estimated functions are step functions and hence not smooth. The BS-lasso achieves large error in all components. Figure 2 shows that the method is often not able to select the four components, and if they are selected, the estimated effects are heavily oscillating and deviating from the true function. The results of the other experiments conducted for Model A are reported in Appendix A and commented below.

First we consider the same scenario with dependent covariates ($t = 1$). The results are reported in Table 4 in Appendix A. Comparing Table 2 and Table 4, we see that when the covariates are not independent, all methods have more difficulties selecting the correct components. Overall, the MS-lasso performs best in terms of true and false positives. For the fourth component both liso and BS-lasso achieve low MSE, while the MS-lasso and the adaptive MS-lasso performs better or comparable for the three other components. However, for this scenario all methods have problems selecting the first two components, and the adaptive liso and BS-lasso tend to select only the fourth component. In Table 5 in the Supplementary Material we observe that when increasing the sample size to $n = 100$ for the dependent covariates, the selection performance is improved for all methods. Our proposed methods are still overall best in selection when considering both true and false positives. As expected we also observe that the MSE’s for each component are smaller for all of the methods compared to the results in Table 4.

The results for the situation with more noise are reported in Table 6 in Appendix A. Obviously the performance of all the methods deteriorates when introducing more noise. The MS-lasso does best in terms of finding the four components in the true model. In terms of estimation, our proposed method is always better or comparable with the linear methods, and always better than B-splines lasso. The adaptive liso is able to select two of the components with highest accuracy, also leading to a smaller MSE for these functions. It seems reasonable that the adaptive liso which provides quite rough estimates of the smooth functions, performs better than the other methods in situations with more noise.

### 3.2 Model B

The results for Model B, with signal to noise ratio $SNR = 4$ and $t = 0$ are reported in Table 7 in Appendix A. The true model has three nonlinear monotone effects, while the last component is strictly linear. In this setting we are interested in investigating the performance if there are both nonlinear monotone effects and linear effects in the model. We should be able to select linear effects by using the MS-lasso, as linear functions are special cases of monotone functions. This is confirmed in the results, and in fact the MS-lasso does just as well, if not better, in terms of false positives as for Model A. Compared to the other methods, the adaptive MS-lasso is overall better in terms of both selection and estimation.

### 3.3 Linear Effects Only

The MS-lasso allows for nonlinear shapes in the estimated functions, which cannot be captured by the linear methods. However, if the effects are really linear, (adaptive) MS-lasso is still able to estimate the true effects. In these experiments we consider the situation when all nonzero effects are linear. We generate $y$ from the linear model: $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$, with $(\beta_1, \beta_2, \beta_3, \beta_4) = (-2, -2, 2, 2)$. The covariates and the noise are generated as before, with $SNR = 4$. Table 3 shows that the (adaptive) MS-lasso is not much worse than the standard lasso and the adaptive lasso which are restricted to linear effects. The linear methods achieve a bit smaller MSE, but in terms of selection, the proposed method is doing just as well. The number of false positives for the MS-lasso...
Figure 1: Estimated functions for (a) the MS-lasso and the adaptive MS-lasso, and (b) the lasso and adaptive lasso. Each line represents one simulation run. The black line is the true curve.
Figure 2: Estimated functions for (a) the BS-lasso and (b) the adaptive liso. Each line represents one simulation run. The black line is the true curve.
and the adaptive MS-lasso is much lower compared to the lasso and the adaptive lasso respectively. The results indicate, however, that both the adaptive liso and B-splines lasso will loose compared to our approach, both in terms of selection and in the ability to fit a straight line. This shows that the possible gain of applying our method is remarkable if the true functions are monotone, but have nonlinear characteristics as in Model A. However, if the effects are truly linear, we are comparable to the methods restricted to linear shapes only. Hence, in general there is much to gain and little to loose using MS-lasso instead of lasso.

4 Data Illustrations

In this section we illustrate the use of our method on two relevant data sets from genomics.

4.1 Bone Mineral Data

For illustration we apply our method to a bone mineral data set previously studied in Reppe et al. (2010). For 84 women who had a trans-iliacal bone biopsy, there are gene expression measurements for 22815 genes. The data were already normalized as described in Reppe et al. (2010) and we fit a regression model with bone mineral density as response and gene expressions as covariates. For simplicity and computational reasons in this illustration, we reduce the dimension to $P = 2000$ covariates which have the larger empirical variance.

The MS-lasso is used to model the relationship between the bone mineral density and the expression of the genes. The covariates are transformed to $[0, 1]$ and we use I-splines of order two with six interior knots. We compare the results with; the lasso and the adaptive lasso for linear effects, the adaptive liso for isotonic effects, as well as the BS-lasso for nonparametric additive models (Huang et al., 2010). For the latter, quadratic B-splines with six evenly distributed knots are used. For all methods we use 10-fold cross-validation to choose the penalty parameters.
Figure 3: Estimated functions for four selected genes from the bone data example.

The six methods select, as expected, different sets of genes. The MS-lasso selects 19 genes, while when including an adaptive step, the adaptive MS-lasso selects 8 genes. The lasso and the adaptive lasso select 14 and 12 variables respectively, while the adaptive liso selects 13 variables. Finally, the BS-lasso selects 8 variables. Out of the 19 genes selected by the MS-lasso, the lasso selects 10 of them and the adaptive lasso 9 of the same genes. The adaptive liso finds 7 of the same genes as the MS-lasso and the BS-lasso finds 5 of the same.

In Figure 3(a) we have plotted the estimated functions for four of the variables that are selected by our method together with the estimated effects by the other methods, if the variable is selected by any of them. Figure 3(a) shows the estimated effects for a variable which is selected by all of the methods. The three methods assuming monotonicity indicate a slow increase for low values of the gene expression, a more rapid increase for midrange values and a plateau effect for high expression values. Obviously, the linear methods are not able to capture this shape, and the BS-lasso is estimating a clear nonmonotone shape which is less reasonable in the current biological application.

In Figure 3(b) we see a similar pattern where the adaptive MS-lasso and the adaptive liso recognize a rapid decrease for midrange values, and it seems that there might be a threshold
Figure 4: Estimated functions for two selected CNVs from the breast cancer illustration.

In cancer studies, recovering relationships between copy number variations (CNV) and gene expression data is of interest as such biological relationships can be relevant for the understanding of disease. Many methods that have been used to investigate these relationships assume linearity, while the underlying biological interaction is not really believed to be linear (Solvang et al., 2011). Solvang et al. (2011) employed a strategy allowing for non-linear dependencies when studying in-cis how a copy number variation would impact the expression. In this example we propose to use MS-lasso to study the dependency between copy number variation and expression also in-trans, assuming monotone nonlinear relationships. Our approach can be used to investigate which CNV’s impact in-trans the expression of a gene G. By defining the response $y$ as the gene expression of gene G and considering the CNV’s as covariates, the aim is to select a set of CNV which describes the variation in gene expression for gene G.

For illustration, we use the combined copy number - gene expression data studied in Solvang et al. (2011) and consider one of the genes, ERBB2, which is discussed in their paper as target variable in the analysis. That is, the gene expression measurements are considered as the response while the copy number variations within the same chromosome as ERBB2 are used as covariates. This becomes a regression problem with $n = 102$ and $P = 668$ covariates.

In this case, the MS-lasso selects six CNV’s. Interestingly, one of them is the copy number corresponding to the gene ERBB2 itself, which is in fact selected by all of the methods. The

4.2 Breast cancer data

In cancer studies, recovering relationships between copy number variations (CNV) and gene expression data is of interest as such biological relationships can be relevant for the understanding of disease. Many methods that have been used to investigate these relationships assume linearity, while the underlying biological interaction is not really believed to be linear (Solvang et al., 2011). Solvang et al. (2011) employed a strategy allowing for non-linear dependencies when studying in-cis how a copy number variation would impact the expression. In this example we propose to use MS-lasso to study the dependency between copy number variation and expression also in-trans, assuming monotone nonlinear relationships. Our approach can be used to investigate which CNV’s impact in-trans the expression of a gene G. By defining the response $y$ as the gene expression of gene G and considering the CNV’s as covariates, the aim is to select a set of CNV which describes the variation in gene expression for gene G.

For illustration, we use the combined copy number - gene expression data studied in Solvang et al. (2011) and consider one of the genes, ERBB2, which is discussed in their paper as target variable in the analysis. That is, the gene expression measurements are considered as the response while the copy number variations within the same chromosome as ERBB2 are used as covariates. This becomes a regression problem with $n = 102$ and $P = 668$ covariates.

In this case, the MS-lasso selects six CNV’s. Interestingly, one of them is the copy number corresponding to the gene ERBB2 itself, which is in fact selected by all of the methods. The
The estimated function is plotted in Figure 4(a). The plot indicates that this effect is actually quite linear, as all methods seem to estimate a shape that is close to linear. However, we also observe that the estimated linear effects might not capture fully the shapes. For example in Figure 4(b) we compare the estimated effects of the MS-lasso and the lasso for another of the selected CNV. From the function estimated by our method, it seems to be a steep decrease with a threshold around 0.4 (in the transformed values of the covariate). The linear estimate is not capturing such a shape and seems to be highly influenced by the few large values in the covariate. The other methods have not selected this CNV.

5 Discussion

The additive model provides a flexible alternative to the standard linear regression model. However, the monotonicity in the linear model is attractive and in some settings it is sensible to preserve monotonicity in the additive models by imposing restrictions on the additive components. In this paper we have proposed a new method for estimation and variable selection in high-dimensional additive models that is restricted to monotone effects. By combining group selection with spline representations of the covariates, the method is within the same category as the methods of for example Meier et al. (2009) and Huang et al. (2010) allowing for nonlinearity in the estimated functions.

We have presented the MS-lasso as a novel and flexible method feasible in high-dimensional settings when $P > n$. Also in low dimensions, if the covariates are represented by a larger number of basis functions or in any other regression setting where variable selection is relevant, our MS-lasso could be useful. The MS-lasso can also be extended to generalized additive models. In this case we simply replace the residual sum of squares in (4) with the relevant (partial) log-likelihood, still representing the covariates by their I-splines basis.

Similar to Meier et al. (2009) or Huang et al. (2010) among others, we fit smooth curves with a fixed number of knots (typically six interior knots), although the placement and number of knots are important issues within the splines literature. Meyer (2008) argues that assumptions on the shape of the function, such as monotonicity, are quite robust to the choice of knots and therefore avoids much of the problems of overfitting the data which are present using other types of (regression) splines. Also, Delecroix and Thomas-Agnan (2000) argue that the number of knots recommended using monotone I-splines is usually much smaller than the number of observation points. Following these arguments, we have not studied the choice of knots in further detail, but this could be a topic of future research.

Another important issue is related to computational efficiency. Though the MS-lasso is feasible when $P > n$, the computational complexity may increase dramatically depending on the number of knots. In situations where the number of inputs to the algorithm becomes too large, it could be reasonable to employ suitable screening or preselection techniques to reduce computational complexity.

We should also address a minor detail about the choice of the penalty parameter $\lambda$. As pointed out in Chiquet et al. (2012), when the cooperative penalty goes to zero, that is for small $\lambda$, sign-coherence is no longer preserved for the cooperative lasso. Hence monotonicity is neither preserved if $\lambda$ is too small. This rarely occurs in our experiments, but to ensure monotonicity we should consider only values of $\lambda$ for which sign-coherence of the cooperative lasso is preserved.

Recently, several approaches doing variable selection in additive partial linear models have been proposed. For example do Wang et al. (2011) and Liu et al. (2011) introduce estimation and variable selection for additive partial linear models with a predefined division of covariates into linear and nonlinear associations with the response. Similarly one may construct models with some covariates strictly linearly related to the response, while a second part of the covariates have nonlinear monotone effect on the response. This is simply achieved by letting the group size in the MS-lasso be equal to one for the linear part of the covariates, indicating that it is straightforward to extend the MS-lasso also to additive partial linear models.

Based on our simulation results and data examples, we conclude that the (adaptive)
Table 4: Results for Model A with $t = 1$ (dependent covariates) and $SNR = 4$: Comparison of the selection and estimation performance for the six methods. To compare the selection performance, the proportion of which the methods select each component in the true model is reported together with the average number of true and false positives. For estimation the mean squared error between the fitted function and the true function, averaged over the 100 simulated data sets, is reported.

| Selection | $g_0^1$ | $g_0^2$ | $g_0^3$ | $g_0^4$ | TP   | FP   |
|-----------|---------|---------|---------|---------|------|------|
| MS-lasso  | 0.16 (0.37) | 0.12 (0.33) | 0.97 (0.17) | 1.00 (0.00) | 2.25 (0.70) | 4.48 (7.25) |
| Ad. MS-lasso | 0.12 (0.33) | 0.11 (0.31) | 0.71 (0.46) | 0.98 (0.14) | 1.92 (0.85) | 0.20 (0.77) |
| Lasso     | 0.00 (0.00) | 0.05 (0.22) | 0.98 (0.14) | 1.00 (0.00) | 2.03 (0.26) | 7.64 (4.01) |
| Ad. lasso | 0.00 (0.00) | 0.05 (0.22) | 0.89 (0.31) | 1.00 (0.00) | 1.94 (0.40) | 5.87 (3.30) |
| Ad. liso  | 0.00 (0.00) | 0.00 (0.00) | 0.04 (0.20) | 1.00 (0.00) | 1.04 (0.20) | 7.10 (4.83) |
| BS-lasso  | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 0.82 (1.71) |

| Estimation | $g_1^0$ | $g_2^0$ | $g_3^0$ | $g_4^0$ |
|------------|---------|---------|---------|---------|
| MS-lasso   | 0.06 (0.01) | 0.16 (0.04) | 0.13 (0.04) | 0.20 (0.05) |
| Ad. MS-lasso | 0.06 (0.02) | 0.15 (0.05) | 0.13 (0.07) | 0.16 (0.08) |
| Lasso      | 0.06 (0.00) | 0.17 (0.01) | 0.16 (0.02) | 0.19 (0.03) |
| Ad. lasso  | 0.06 (0.00) | 0.16 (0.03) | 0.14 (0.03) | 0.16 (0.04) |
| Ad. liso   | 0.06 (0.00) | 0.17 (0.00) | 0.21 (0.01) | 0.14 (0.05) |
| BS-lasso   | 0.06 (0.00) | 0.17 (0.00) | 0.21 (0.00) | 0.12 (0.04) |

MS-lasso shows excellent properties compared to other methods and should be advised for high-dimensional monotone regression.

Acknowledgements
We thank Hiroko Solvang for providing the breast cancer data, and Sjur Reppe who provided the bone biopsy data.

Appendix A
This section contains four tables with results from simulation experiments. The first two tables, Table 4 and Table 5 report the results from experiments using Model A, as described in the paper, and dependent covariates. We consider sample size $n = 50$ in Table 4, and $n = 100$ in Table 5. In Table 6, results from Model A with more noise are reported. Finally, Table 7 contains the results from Model B. All experiments are described and discussed in Section 3.
Table 5: Results for Model A with $t = 1$ (dependent covariates), $SNR = 4$ and $n = 100$: Comparison of the selection and estimation performance for the six methods. To compare the selection performance, the proportion of which the methods select each component in the true model is reported together with the average number of true and false positives. For estimation the mean squared error between the fitted function and the true function, averaged over the 100 simulated data sets, is reported.

| Selection | $g_0^0$ | $g_1^0$ | $g_2^0$ | $g_3^0$ | $g_4^0$ | TP      | FP      |
|-----------|---------|---------|---------|---------|---------|---------|---------|
| MS-lasso  | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 4.00 (0.00) | 1.53 (1.41) |
| Ad. MS-lasso | 0.78 (0.42) | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 3.78 (0.42) | 0.00 (0.00) |
| Lasso     | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 4.00 (0.00) | 23.16 (16.95) |
| Ad. lasso | 0.71 (0.46) | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 3.71 (0.46) | 3.27 (2.53) |
| BS-lasso  | 0.14 (0.35) | 0.99 (0.10) | 1.00 (0.00) | 1.00 (0.00) | 3.13 (0.37) | 5.81 (5.09) |

Table 6: Results for Model A with $SNR = 2$: Comparison of the selection and estimation performance for the six methods. To compare the selection performance, the proportion of which the methods select each component in the true model is reported together with the average number of true and false positives. For estimation the mean squared error between the fitted function and the true function, averaged over the 100 simulated data sets, is reported.

| Selection | $g_0^0$ | $g_1^0$ | $g_2^0$ | $g_3^0$ | $g_4^0$ | TP      | FP      |
|-----------|---------|---------|---------|---------|---------|---------|---------|
| MS-lasso  | 0.50 (0.50) | 0.25 (0.44) | 0.93 (0.26) | 0.89 (0.31) | 2.57 (1.04) | 11.74 (9.92) |
| Ad. MS-lasso | 0.33 (0.47) | 0.18 (0.39) | 0.92 (0.27) | 0.73 (0.45) | 2.16 (1.05) | 3.64 (4.27) |
| Lasso     | 0.37 (0.49) | 0.25 (0.44) | 0.85 (0.36) | 0.86 (0.35) | 2.33 (1.12) | 19.54 (14.19) |
| Ad. lasso | 0.30 (0.46) | 0.20 (0.40) | 0.85 (0.36) | 0.83 (0.38) | 2.18 (1.05) | 15.15 (10.43) |
| BS-lasso  | 0.11 (0.31) | 0.45 (0.50) | 0.83 (0.38) | 0.92 (0.27) | 2.31 (1.03) | 7.73 (4.87) |

| Estimation | $g_0^0$ | $g_1^0$ | $g_2^0$ | $g_3^0$ | $g_4^0$ | TP      | FP      |
|-----------|---------|---------|---------|---------|---------|---------|---------|
| MS-lasso  | 0.13 (0.05) | 0.30 (0.05) | 0.35 (0.15) | 0.34 (0.15) |
| Ad. MS-lasso | 0.12 (0.06) | 0.29 (0.09) | 0.18 (0.18) | 0.27 (0.21) |
| Lasso     | 0.14 (0.03) | 0.30 (0.05) | 0.48 (0.15) | 0.34 (0.16) |
| Ad. lasso | 0.13 (0.05) | 0.29 (0.07) | 0.40 (0.18) | 0.27 (0.19) |
| BS-lasso  | 0.16 (0.01) | 0.32 (0.02) | 0.69 (0.14) | 0.35 (0.23) |
Table 7: Results for Model B with SNR = 4: Comparison of the selection and estimation performance for the six methods. To compare the selection performance, the proportion of which the methods select each component in the true model is reported together with the average number of true and false positives. For estimation the mean squared error between the fitted function and the true function, averaged over the 100 simulated data sets, is reported.

|       | $g_1^0$  | $g_2^0$  | $g_3^0$  | $g_4^0$  | TP       | FP       |
|-------|----------|----------|----------|----------|----------|----------|
| MS-lasso | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00) | 4.00 (0.00) | 22.27 (8.64) |
| Ad. MS-lasso | 0.97 (0.17) | 0.98 (0.14) | 1.00 (0.00) | 0.98 (0.14) | 3.93 (0.43) | 1.41 (2.22) |
| Lasso | 0.93 (0.26) | 0.96 (0.20) | 1.00 (0.00) | 1.00 (0.00) | 3.89 (0.40) | 31.33 (11.32) |
| Ad. lasso | 0.85 (0.36) | 0.93 (0.26) | 1.00 (0.00) | 1.00 (0.00) | 3.78 (0.52) | 21.39 (6.87) |
| Ad. liso | 0.35 (0.48) | 0.95 (0.22) | 1.00 (0.00) | 0.91 (0.29) | 3.21 (0.67) | 6.95 (3.33) |
| BS-lasso | 0.01 (0.10) | 0.08 (0.27) | 0.55 (0.50) | 0.15 (0.36) | 0.79 (0.86) | 3.77 (4.13) |

|       | $g_1^0$  | $g_2^0$  | $g_3^0$  | $g_4^0$  |
|-------|----------|----------|----------|----------|
| MS-lasso | 0.04 (0.02) | 0.10 (0.03) | 0.10 (0.03) | 0.06 (0.03) |
| Ad. MS-lasso | 0.02 (0.03) | 0.04 (0.05) | 0.03 (0.04) | 0.02 (0.05) |
| Lasso | 0.10 (0.03) | 0.19 (0.06) | 0.27 (0.06) | 0.11 (0.06) |
| Ad. lasso | 0.09 (0.04) | 0.15 (0.07) | 0.23 (0.04) | 0.08 (0.06) |
| Ad. liso | 0.13 (0.05) | 0.09 (0.07) | 0.09 (0.05) | 0.11 (0.09) |
| BS-lasso | 0.16 (0.01) | 0.32 (0.03) | 0.52 (0.22) | 0.31 (0.06) |

References

Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association*, 84(405):289–294.

Barlow, R., Bartholomew, D., Brenner, J., and Brunk, H. (1972). *Statistical inference under order restrictions; the theory and application of isotonic regression*. New York: Wiley.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Verlag Berlin Heidelberg. Springer Series in Statistics. Springer.

Chiquet, J., Grandvalet, Y., and Charbonnier, C. (2012). Sparsity with sign-coherent groups of variables via the cooperative-lasso. *Annals of Applied Statistics*, 6(2):795–830.

Delecroix, M. and Thomas-Agnan, C. (2000). Spline and kernel regression under shape restrictions. In *Smoothing and Regression: Approaches, computations and applications*. Hoboken, NJ, USA. Wiley-Interscience Publication.

Fang, Z. and Meinshausen, N. (2012). Lasso isotone for high-dimensional additive isotonic regression. *Journal of Computational and Graphical Statistics*, 21(2):72–91.

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

He, X. and Shi, P. (1998). Monotone B-spline smoothing. *Journal of the American Statistical Association*, 93:643–650.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38:2282–2313.

Liu, X., Wang, L., and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21(3):1225–1248.
Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, 52(1):374–393.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *Annals of Applied Statistics*, 2(3):1013–1033.

Raftery, A. E. and Richardson, S. (1996). Model selection for generalized linear models via glib: Application to nutrition and breast cancer. In *Bayesian Biostatistics*. New York. Marcel Dekker, Inc.

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71(5):1009–1030.

Reppe, S., Refvem, H., Gautvik, V. T., Olstad, O. K., Hvring, P. I., Reinholt, F. P., Holden, M., Frigessi, A., Jentland, R., and Gautvik, K. M. (2010). Eight genes are highly associated with BMD variation in postmenopausal caucasian women. *Bone*, 46(3):604–612.

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order restricted statistical inference*. New York: Wiley.

Solvang, H. K., Lingjærde, O., Frigessi, A., Barresen-Dale, A.-L., and Kristensen, V. N. (2011). Linear and non-linear dependencies between copy number aberrations and mrna expression reveal distinct molecular pathways in breast cancer. *BMC Bioinformatics*, 12:197.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288.

Tibshirani, R. J., Hothorn, H., and Tibshirani, R. (2011). Nearly-isotonic regression. *Technometrics*, 53(1):54–61.

Tutz, G. and Leitenstorfer, F. (2007). Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics*, 16(1):165–188.

van de Geer, S., Bühlmann, P., and Zhou, S. (2011). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749.

Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Annals of Statistics*, 39(4):1827–1851.

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

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.