Tutz, Ulbricht:

Penalized Regression with Correlation Based Penalty

Penalized Regression with Correlation Based Penalty

Gerhard Tutz & Jan Ulbricht
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
{tutz,ulbricht}@stat.uni-muenchen.de

15th September 2006

Abstract

A new regularization method for regression models is proposed. The criterion to be minimized contains a penalty term which explicitly links strength of penalization to the correlation between predictors. As the elastic net, the method encourages a grouping effect where strongly correlated predictors tend to be in or out of the model together. A boosted version of the penalized estimator, which is based on a new boosting method, allows to select variables. Real world data and simulations show that the method compares well to competing regularization techniques. In settings where the number of predictors is smaller than the number of observations it frequently performs better than competitors, in high dimensional settings prediction measures favor the elastic net while accuracy of estimation and stability of variable selection favors the newly proposed method.

Keywords: Correlation based estimator, Boosting, Variable selection, Elastic net, Lasso, Penalization.

1 Introduction

We focus on the usual linear regression model

$$y = \beta_0 + \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$$

where $\mathbf{x}^T = (x_1, \ldots, x_p)$ is a vector of covariates and ε is a noise variable with $E(\varepsilon) = 0$. In particular for high dimensional predictors $\mathbf{x}$, the ordinary least squares (OLS) estimate may not be unique. Moreover, it is not the first choice if the aim is prediction. Alternative estimators like the ridge regression estimator (Hoerl & Kennard (1970)) do much better and are unique for an appropriately chosen shrinkage parameter.

Within the last decade many alternative shrinkage estimators have been proposed, in particular the lasso (Tibshirani (1996)) which imposes an L_1-penalty on the regression coefficients. By using a non-convex penalty it does automatic variable selection in contrast to the ridge regression estimator which only shrinks the estimates towards zero. More recently, Zou & Hastie (2005) proposed the elastic net as an alternative procedure which handles deficiencies of lasso and ridge regression by combining the L_2 and L_1 penalty. One motivation Zou & Hastie (2005) give for the elastic net is its property to include groups of variables which are highly correlated. If variables are highly correlated, as for example gene expression in microarray data, the lasso selects only one of the group whereas the elastic net catches "all the big fish", meaning that it selects the whole group.

In this paper an alternative regularization procedure is proposed which aims at the selection of groups of correlated variables. In the simpler version it is based on a penalty that explicitly uses correlation between variables as weights. In the extended version boosting techniques are used for groups of variables.

2 Penalized regression linked to correlation

Let the data be given by $(y_i, \mathbf{x_i}), i = 1, \ldots, n$, with y_i denoting the response and $\mathbf{x_i}^T = (x_{i1}, \ldots, x_{ip})$ denoting the predictor. For simplicity the response and the covariates are considered as centered. Regularized estimates of the parameter vector $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ may be obtained by minimizing the penalized least squares criterion

$$PLS = \sum_{i=1}^{n} |y_i - \mathbf{x_i}^T \boldsymbol{\beta}|^2 + P(\boldsymbol{\beta}) \tag{1}$$

where $P(\boldsymbol{\beta})$ is a specific penalty term. Common penalties are of the bridge penalty type (Frank & Friedman (1993), Fu (1998))

$$P(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{P} |\beta_j|^{\gamma}, \gamma > 0,$$

where λ is a tuning parameter. For $\gamma = 2$ one obtains ridge regression (Hoerl & Kennard (1970)), for $\gamma = 1$ the lasso (Tibshirani (1996)). Penalties with $\gamma < 1$ have also been called soft thresholding (Donoho & Johnstone (1995), Klinger (1998)). The more recently proposed elastic net (Zou & Hastie (2005)) is based on a combination of the ridge penalty and the lasso by using a penalty term with two tuning parameters λ_1, λ_2 given by

$$P(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{P} |\beta_j| + \lambda_2 \sum_{j=1}^{P} \beta_j^2.$$

The method inherits properties of the lasso by doing variable selection, but in situations where ridge performs better ($n > p$ and high correlation between predictors) it relies on the ridge type penalty. The elastic net tends to include highly correlated predictors rather then selecting one of them.

2.1 The correlation based estimator

The method proposed here utilizes the correlation between predictors explicitly in the penalty term. Coefficients which correspond to pairs of covariates are weighted according to their marginal correlation. The *correlation based penalty* is given by

$$\begin{aligned} P_c(\boldsymbol{\beta}) &= \lambda \sum_{i=1}^{p-1} \sum_{j>i} \left\{ \frac{(\beta_i - \beta_j)^2}{1 - \varrho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1 + \varrho_{ij}} \right\} \\ &= 2\lambda \sum_{i=1}^{p-1} \sum_{j>i} \frac{\beta_i^2 - 2\varrho_{ij}\beta_i\beta_j + \beta_j^2}{1 - \varrho_{ij}^2} \end{aligned} \quad (2)$$

where ϱ_{ij} denotes the (empirical) correlation between the ith and the jth predictor. It is designed in a way so that for strong positive correlation ($\varrho_{ij} \to 1$) the first term becomes dominant having the effect that estimates for β_i, β_j are similar ($\hat{\beta}_i \approx \hat{\beta}_j$). For strong negative correlation ($\varrho_{ij} \to -1$) the second term becomes dominant and $\hat{\beta}_i$ will be close to $-\hat{\beta}_j$. The effect is grouping, highly correlated effects show comparable values of estimates ($|\hat{\beta}_i| \approx |\hat{\beta}_j|$) with the sign being determined by positive or negative correlation. If the predictors are uncorrelated ($\varrho_{ij} = 0$) one obtains (up to a constant) the ridge penalty $P_c(\boldsymbol{\beta}) \propto \lambda \sum \beta_i^2$. Consequently, for weakly correlated data the performance is quite close to the ridge regression estimator. Therefore, as in the elastic net ridge regression is a limiting case.

A nice feature of the penalty (2) is that it may be given as a simple quadratic form which allows to give the resulting estimator in closed form. One obtains

$$P_c(\boldsymbol{\beta}) = \lambda \boldsymbol{\beta}^T \mathbf{W} \boldsymbol{\beta}$$

where $\mathbf{W}$ is a matrix that is determined by the correlations $\varrho_{ij}, i, j = 1, \ldots, p$ (for details on $\mathbf{W}$ see next section). For $\varrho_{ij}^2 \neq 1, \lambda > 0$, an explicit solution to the penalized least squares problem (1) is obtained by the *correlation based estimator*

$$\hat{\boldsymbol{\beta}}_c = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{W})^{-1} \mathbf{X}^T \mathbf{y}, \quad (3)$$

where $\mathbf{X}^T = (\mathbf{x_1}, \ldots, \mathbf{x_n})$ is the design matrix and $\mathbf{y}$ collects the responses, $\mathbf{y}^T = (y_1, \ldots, y_n)$.

2.2 Structure of the penalty

The grouping effect strongly depends on the convexity of the penalty term. The correlation based penalty may be seen as a combination of two penalties, $P_c(\boldsymbol{\beta}) = P_{c,1}(\boldsymbol{\beta}) + P_{c,2}(\boldsymbol{\beta})$ where

$$\begin{aligned} P_{c,1}(\boldsymbol{\beta}) &= \lambda \sum_i \sum_{j>i} \frac{(\beta_i - \beta_j)^2}{1 - \varrho_{ij}}, \\ P_{c,2}(\boldsymbol{\beta}) &= \lambda \sum_i \sum_{j>i} \frac{(\beta_i + \beta_j)^2}{1 + \varrho_{ij}}. \end{aligned}$$

The first term becomes influential for positively correlated predictors whereas the latter term is influential for negatively correlated predictors. Neither $P_{c,1}(.)$ nor $P_{c,2}(.)$ is strictly convex. But (for $\lambda > 0$ and $\varrho_{ij}^2 \neq 1$ if $i \neq j$) the combination $P_c(\beta)$ is strictly convex. A nice consequence is that the estimate $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ exists and is unique.

Proposition 1:
Assume that $\lambda > 0$ and $\varrho_{ij}^2 \neq 1$ for $i \neq j$. Then one obtains

(1) $P_c(\boldsymbol{\beta})$ is strictly convex.

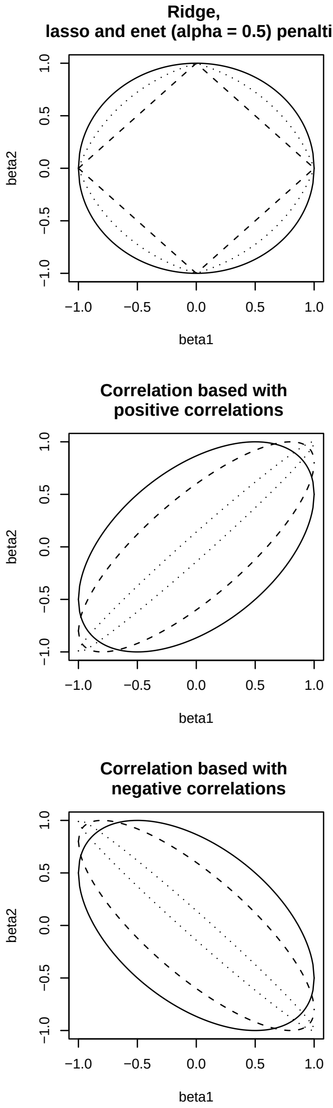
(2) The estimate $\hat{\boldsymbol{\beta}}_c$ exists and is unique.

(3) $P_c(\boldsymbol{\beta})$ may be given as a quadratic form $P_c(\boldsymbol{\beta}) = \lambda \boldsymbol{\beta}^T \mathbf{W} \boldsymbol{\beta}$ where $\mathbf{W} = (w_{ij})$ is determined by

$$w_{ij} = \begin{cases} 2\sum_{s \neq i} \frac{1}{1-\varrho_{is}^2}, & i = j, \\ -2\frac{\varrho_{ij}}{1-\varrho_{ij}^2}, & i \neq j \end{cases} \tag{4}$$

(for proof see Appendix). Thus for $\lambda > 0$ the correlation based estimate shares the property of existence and uniqueness with the ridge estimator. In contrast, the lasso estimate does not necessarily have a unique solution.

Figure 1 shows the two-dimensional contour plots for selected values of ϱ. The constraint region for the ridge penalty is the disk $\beta_1^2 + \beta_2^2 \leq c$, for the Lasso one obtains the diamond $|\beta_1| + |\beta_2| \leq c$. Since the diamond has distinct corners, if a solution occurs at a corner then one parameter β_j is equal to zero. It is seen that contours for ridge and Lasso are highly symmetric, $x_1 = 0$ is an axis of symmetry as well as $x_2 = 0$. In contrast, the constrained region for the correlation based estimator is an ellipsoid which becomes narrower with increasing



Ridge,
lasso and enet (alpha = 0.5) penalties

Correlation based with
positive correlations

Correlation based with
negative correlations

Figure 1: *Top panel: Two-dimensional contour plots of penalized least squares with $\varrho = 0$ (solid line), LASSO penalty (dashed line) and elastic net with $\alpha = 0.5$ (dotted line). Mid panel: Two-dimensional contour plots of correlation based penalty for three amounts of positive correlation: $\varrho = 0.5$ (solid line), $\varrho = 0.8$ (dashed line), and $\varrho = 0.99$ (dotted line). Bottom panel: Two-dimensional contour plots of correlation based penalty for three amounts of negative correlation: $\varrho = -0.5$ (solid line), $\varrho = -0.8$ (dashed line), and $\varrho = -0.99$ (dotted line)*

correlation. Spectral decomposition of $P_c(\beta)$ yields eigenvectors $(1, 1)$ and $(1, -1)$ with corresponding eigenvalues $\lambda/(1-\varrho)$ and $\lambda/(1+\varrho)$. Thus for $\varrho > 0$ the first eigenvalue becomes dominant while for $\varrho < 0$ it is the second eigenvalue that determines the orientation of the ellipsoid. When computing the penalized least squares criterion the effect is that for $\varrho > 0$ estimates are preferred for which the components $\hat{\beta}_1, \hat{\beta}_2$ are similar, for $\varrho < 0$ similarity of $\hat{\beta}_1$ and $-\hat{\beta}_2$ is preferred. This may be seen from the contour plots, since for $\varrho > 0$ the increase in $P_c(\beta)$ is slower when moving into the direction of the first eigenvector $(1, 1)$ than into the orthogonal direction $(1, -1)$. For $\varrho < 0$ the eigenvalue corresponding to $(1, -1)$ is larger and therefore parameter values where β_1 is close to $-\beta_2$ are preferred. Thus the use of penalty P_c implies shrinkage with the strength of shrinkage being determined by λ, but shrinkage differs from ridge shrinkage which occurs for the special case $\varrho_{ij} = 0$.

2.3 Grouping effect: the extreme case

A regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal (up to a change of sign). For the generic penalization method (1) it has been shown that for identical covariate vectors $x_i = x_j$ one obtains $\hat{\beta}_i = \hat{\beta}_j$, if $P(\boldsymbol{\beta})$ is strictly convex (see Lemma 2 of Zou & Hastie (2005)). However, for the correlation based estimator

$$\hat{\boldsymbol{\beta}}_c = \arg\min_{\boldsymbol{\beta}} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + P_c(\boldsymbol{\beta}) \tag{5}$$

the explicit solution $\hat{\boldsymbol{\beta}}_c = (\mathbf{X}^T\mathbf{X} + \mathbf{P}_c(\boldsymbol{\beta}))^{-1}\mathbf{X}^T\mathbf{y}$ is available only for not perfectly correlated predictors. If $x_i = x_j$ the correlation based penalty is no longer strictly convex and Lemma 2 of Zou & Hastie (2005) does not apply. However, although for $\varrho_{ij}^2 \to 1$ the penalty $P_c(\boldsymbol{\beta})$ deteriorates, the estimate may be defined as the limit. With $\beta_c(\lambda, \{\varrho_{ij}\})$ denoting the solution of (5) one defines for $\varrho_{ij}^2 = 1$ the correlation based estimator by

$$\hat{\boldsymbol{\beta}}_c(\lambda, \{\varrho_{ij}\}) = \lim_{\tilde{\varrho}_{ij}^2 \to 1} \boldsymbol{\beta}_c(\lambda, \{\tilde{\varrho}_{ij}\})$$

where the limit is taken for $\tilde{\varrho}_{ij} \to 1$ if $x_i = x_j$ and $\tilde{\varrho}_{ij} \to -1$ if $x_i = -x_j$. For all practical purposes we found $\tilde{\varrho}_{ij} = 0.98$ to work well as a substitute for the limit estimate. For illustration the special case $p = 2$ is considered more closely. One obtains $P_c(\boldsymbol{\beta}) = \lambda\boldsymbol{\beta}^T\mathbf{W}\boldsymbol{\beta} = \lambda\boldsymbol{\beta}^{\mathbf{T}}\mathbf{D_2^T}\mathbf{D_2}\boldsymbol{\beta}$ where

$$\begin{aligned}
\mathbf{D_2} &= \begin{pmatrix} 1/\sqrt{1-\varrho} & -1/\sqrt{1-\varrho} \\ 1/\sqrt{1+\varrho} & 1/\sqrt{1+\varrho} \end{pmatrix}, \\
\mathbf{W} &= \frac{2}{1-\varrho^2}\begin{pmatrix} 1 & -\varrho \\ -\varrho & 1 \end{pmatrix}.
\end{aligned}$$

For the limiting case $\varrho_2^2 \to 1$ the inverse may be computed explicitly. One obtains

$$\begin{aligned}\lim_{\varrho\to 1}(\mathbf{X^TX}+\lambda\mathbf{W})^{\mathbf{-1}} &= \frac{1}{2(2+\lambda)}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix},\\ \lim_{\varrho\to -1}(\mathbf{X^TX}+\lambda\mathbf{W})^{\mathbf{-1}} &= \frac{1}{2\lambda}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}.\end{aligned}$$

It is directly seen that in the limit one obtains $\hat{\beta}_1 = \hat{\beta}_2$ for $\varrho = 1$ and $\hat{\beta}_1 = -\hat{\beta}_2$ for $\varrho = -1$.
Moreover, in the special case $p = 2$ the eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{W}$ are the same and it may be derived that, given the ordinary least squares estimate $\hat{\boldsymbol{\beta}}_{OLS}$ exists, the correlation based estimator is a shrinked version $\hat{\boldsymbol{\beta}}_c = \gamma\hat{\boldsymbol{\beta}}_{OLS}$ where $\gamma = (1-\varrho^2)/(1-\varrho^2+2\lambda)$. If $\varrho \neq 0$ this is different from ridge regression where shrinkage is with respect to the orthonormal basis which spans the column space of $\mathbf{X}$ (e.g. Hastie, Tibshirani & Friedman (2001), Section 3).

3 Simulations in medium-dimensional settings

In the following we first investigate the performance of several methods for a medium number of variables. The simulation setting is similar to the setting used in the original lasso paper (Tibshirani (1996)) and the elastic net paper (Zou & Hastie (2005)). The underlying regression model is given by

$$y = \mathbf{x}^\top\boldsymbol{\beta} + \sigma\varepsilon, \quad \varepsilon \sim N(0,1).$$

Each data set consists of a training set, on which the model were fitted, a validation set, which was used to select the tuning parameters, and a test set for evaluation of the performance. The notation $\cdot|\cdot|\cdot$ is used to describe the number of observations in the training, validation and test set, respectively. In simulations, we center all variables based on the training data set. Let $\bar{\mathbf{x}}_{train}^\top = (\bar{x}_{1,train},\ldots,\bar{x}_{p,train})$ denote the vector of means of the training data, n_{test} the number of observations in the test data set and $\bar{y}_{train}$ the mean over responses in the training data.

We use two measures of performance, the test error (mean squared error) $MSE_y = \frac{1}{n_{test}}\mathbf{r}_{sim}^\top\mathbf{r}_{sim}$ with $r_{i,sim} = \mathbf{x}_i^\top\boldsymbol{\beta} - (\bar{y}_{train} + (\mathbf{x}_i - \bar{\mathbf{x}}_{train})^\top\hat{\boldsymbol{\beta}})$ on the test data set and the mean squared error for the estimation of $\boldsymbol{\beta}$, $MSE_\beta = |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|^2$. While the first measure evaluates prediction, the latter aims at the accuracy of the estimator and therefore the identification of the effect strength of variables.

The scenarios which have been investigated by simulating 50 data sets are given in the following.

(1) In the first example with $p = 8$, $\boldsymbol{\beta}$ is specified by $\boldsymbol{\beta}^\top = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. The pairwise correlation was set to $\varrho(x_i, x_j) = 0.5^{|i-j|}$. The sample size was 20|20|200.

(2) With $p = 9$, $\boldsymbol{\beta}$ is specified by $\boldsymbol{\beta}^\top = (1,2,3,4,0,4,3,2,1)$ and $\sigma = 3$, $\varrho(x_i, x_j) = 1 - 0.25|i-j|$, same sample size as in (1).

(3) The setting is the same as in (1) except that $\beta_1 = \beta_2 = \ldots = \beta_8 = 0.85$.

(4) With $p = 40$, the coefficient vector is given by

$$\boldsymbol{\beta}^\top = (\underbrace{0,\ldots,0}_{10}, \underbrace{2,\ldots,2}_{10}, \underbrace{0,\ldots,0}_{10}, \underbrace{2,\ldots,2}_{10}),$$

$\sigma = 15$, $\varrho(x_i, x_j) = 0.5$, for all i and j. The sample size was 100|100|400.

Method	Simulation 1		Simulation 2		Simulation 3		Simulation 4	
	median MSE_y	median MSE_β	median MSE_y	median MSE_β	median MSE_y	median MSE_β	median MSE_y	median MSE_β
Ridge Regr.	3.28	3.35	3.57	17.12	**1.91**	1.69	30.25	51.67
LASSO	**2.92**	**3.13**	3.73	28.83	3.35	3.99	43.69	83.26
Elastic Net	2.96	3.65	4.15	23.91	3.46	4.44	47.95	87.64
CP	3.40	3.68	**3.06**	**16.95**	2.07	**1.20**	**21.95**	**34.47**

TABLE 1: *Median test mean squared errors and median MSE_β for the simulated examples 1-4, based on 50 replications.*

The simulation results are given in Table 1 and Figure 2. In Table 1 the best performance is given in boldface. The first example contains only positively correlated variables whereas in the second example variables are positively and negatively correlated. Examples 3 and 4 contain grouped variables; in example 3 there is only one group, in example 4 there are two groups of relevant variables. Examples 1, 3, and 4 correspond to examples 1, 2, and 3 in Zou & Hastie (2005). It is seen that in the first setting with positively correlated variables the elastic net performs best but the performance does not strongly differ between methods. In the case of positively and negatively correlated variables the correlation based estimate dominates. In particular in the last setting with grouping effect the correlation based estimator clearly dominates all of the other methods in both measures of performance.

4 Blockwise Boosting

The correlation based estimator (2) does shrinkage but not variable selection. Thus, in particular for very high dimensional predictors it has some drawbacks. A method that performs very well in high dimensions is componentwise boosting.

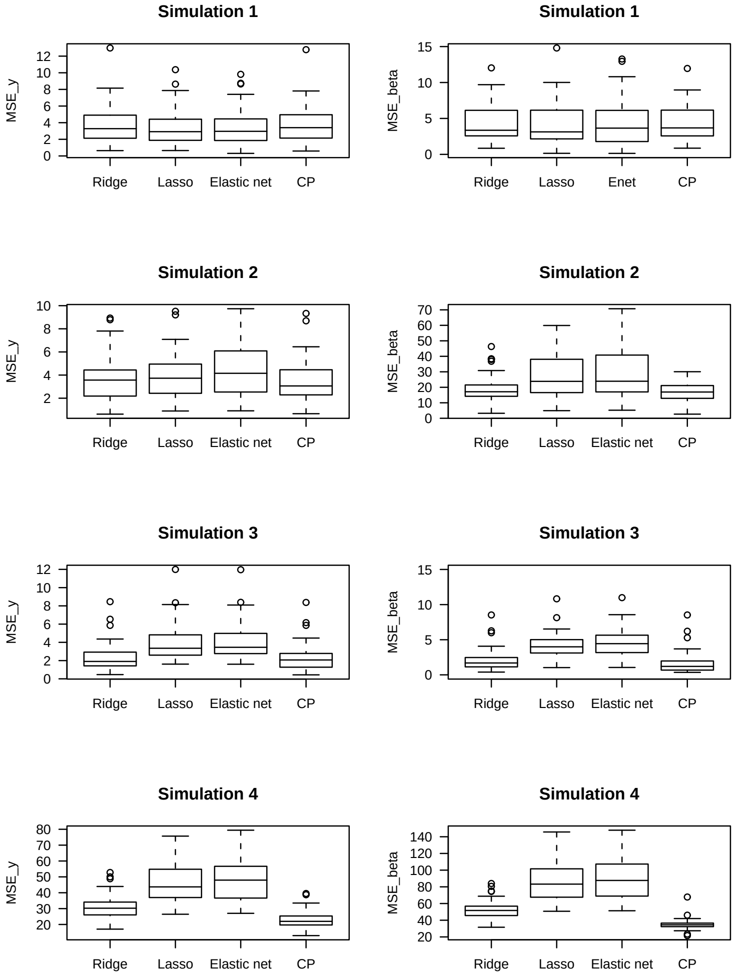


FIGURE 2: *Boxplots of test mean-squared errors (left column) and MSE_β (right column) for simulations 1-4.*

Originating in the machine learning community it has been shown to have good properties in regression (Bühlmann & Yu (2003), Bühlmann (2006)).

To overcome the drawbacks of the correlation based estimator we propose a novel way of componentwise boosting. In order to obtain the grouping effect of the correlation based estimator in combination with variable selection we consider a boosting procedure which updates in each step the coefficients of more than one variable. The procedure differs from common componentwise boosting where just one variable is selected and the corresponding coefficient is adjusted. To distinguish componentwise boosting from the procedure considered here we will refer to blockwise boosting.

Let $S \subset \{1, \ldots, p\}$ denote the index set of the variables which are considered in a given step. The basic concept is to compute within one step of the iterative procedure the parameters which minimize the penalized least squares criterion

$$|\mathbf{r} - \mathbf{X}_S \mathbf{b}|^2 + P_{c,S}, \qquad (6)$$

where $\mathbf{r}$ denotes the vector of residuals (from the previous step), $\mathbf{X}_S$ is the reduced design matrix containing only the variables $j \in S$ and $P_{c,S}$ is the correlation based penalty for the subset S. By minimization of (6) one obtains a simultaneous refit for all the components of S. As usual in boosting, in each step a weak learner is used. That means only a small change in parameter estimates should occur within one step. Thus the parameter λ in (6) is chosen very large, in our case $\lambda \geq 1000$. It has been shown (Bühlmann & Yu (2003)) that large values of λ yield better performances. The only limit is the computational effort, since very large values of λ entail many iteration steps.

As in componentwise boosting, variable selection is performed by selecting in each step an appropriate subset S. Considering all possible subsets implies heavy computational effort, even for a small number of variables. Therefore the candidate sets are reduced by first ordering the variables (as in componentwise boosting) and then considering as candidate sets only subsets S that are built by successively adding one variable from the given order. Therefore in each step first an ordering of variables is constructed.

For subsets S that contain only one variable P_c from (2) cannot be used directly. In those cases we define the penalty by the ridge type penalty $P_{c,\{j\}} = \lambda \beta_j^2$. The adequacy of subsets is evaluated by information theory measures as the AIC, which is also used as a stopping criterion. First we will give the algorithm, deferring the specification of the used AIC criterion until later.

Algorithm BlockBoost

Step 1: (Initialization)

Set $\hat{\boldsymbol{\beta}}^{(0)} = 0, \hat{\boldsymbol{\mu}}^{(0)} = 0$.

Step 2: (Iteration)

For $m = 1, 2, \ldots$

(a) *Find an appropriate order of regressors according to their improvements of fit*

Compute the residuals $\mathbf{r}^{(m)} = \mathbf{y} - \hat{\boldsymbol{\mu}}^{(m-1)}$ and fit for $j \in \{1, \ldots, p\}$ the model $\mathbf{r}^{(m)} = \mathbf{X}_{\{j\}} b_j + \boldsymbol{\epsilon}$ by minimizing $\mid \mathbf{r}^{(m)} - \mathbf{X}_{\{j\}} b_j \mid^2 + P_{c,\{j\}}$, yielding $\hat{b}_{j_1}, \ldots, \hat{b}_{j_p}$ such that $AIC(\hat{b}_{j_1}) \leq \ldots \leq AIC(\hat{b}_{j_p})$.

(b) *Find a suitable number of regressors to update*

For $r = 1, \dots, p$

With $S_r = \{j_1, \dots, j_r\}$ fit the model $\mathbf{r}^{(m)} = \mathbf{X}_{S_r}\mathbf{b}_{S_r} + \boldsymbol{\epsilon}$ by minimizing $| \mathbf{r}^{(m)} - \mathbf{X}_{S_r}\mathbf{b}_{S_r} |^2 + P_{c,S_r}$ yielding estimates $\hat{\mathbf{b}}_{S_r}$ and AIC criterion $AIC(\hat{\mathbf{b}}_{S_r})$.

(c) *Selection*

Select the subset of variables which has the best fit, yielding

$$S^{(m)} = \arg\min_{S_r} AIC(\hat{\mathbf{b}}_{S_r}).$$

(d) *Refit*

The parameter vector is updated by

$$\hat{\beta}_j^{(m)} = \begin{cases} \hat{\beta}_j^{(m-1)} + \hat{b}_j, & \text{if } j \in S^{(m)}, \\ \hat{\beta}_j^{(m-1)}, & \text{otherwise,} \end{cases}$$

yielding the vector $\hat{\boldsymbol{\beta}}^{(m)} = (\hat{\beta}_1^{(m)}, \dots, \hat{\beta}_p^{(m)})^T$ and $\hat{\boldsymbol{\mu}}^{(m)} = \hat{\boldsymbol{\mu}}^{(m-1)} + \mathbf{X}_{S^{(m)}}\hat{\mathbf{b}}_{S^{(m)}}$.

The stopping criterion we propose is a version of the AIC criterion $AIC = -2(l(\hat{\boldsymbol{\mu}}^{(m)}) - tr(\mathbf{H}_m))$ where $l(\hat{\boldsymbol{\mu}}^{(m)})$ denotes the log-likelihood after the mth refit and $tr(\mathbf{H}_m)$ is the trace of the corresponding hat matrix. Some derivation shows that it is given by

$$\widehat{\boldsymbol{\mu}}^{(m)} = \mathbf{H}_m \mathbf{y},$$

where

$$\begin{aligned} \mathbf{H}_m &= \sum_{j=1}^{m} \tilde{\mathbf{H}}_j \prod_{\ell=1}^{j-1} (I - \tilde{\mathbf{H}}_{j-\ell}) \\ &= \tilde{\mathbf{H}}_1 + \tilde{\mathbf{H}}_2(I - \tilde{\mathbf{H}}_1) + \dots \end{aligned}$$

with $\tilde{\mathbf{H}}_j = \mathbf{X}_{S^{(j)}}(\mathbf{X}_{S^{(j)}}^\top \mathbf{X}_{S^{(j)}} + \lambda \mathbf{W}_{S^{(j)}})^{-1}\mathbf{X}_{S^{(j)}}^\top$, where $\mathbf{W}_{S^{(j)}}$ denotes the penalty matrix from (3) for subset $S^{(j)}$.

We use the corrected AIC criterion (Hurvich, Simonoff & Tsai (1998)) with an additional correction factor

$$AIC_c = \log(\widehat{\sigma}_m^2) + \frac{1 + 1.8 \cdot tr(\mathbf{H}_m)/n}{1 - (1.8 \cdot tr(\mathbf{H}_m) + 2)/n},$$

where

$$\widehat{\sigma}_m^2 = \frac{1}{n}(\mathbf{y} - \widehat{\boldsymbol{\mu}}^{(m)})^\top (\mathbf{y} - \widehat{\boldsymbol{\mu}}^{(m)}).$$

The estimate $\hat{\boldsymbol{\beta}}_B$ resulting from BlockBoost inherits the strong grouping effect from the correlation based estimator. If predictors are highly correlated the corresponding updates within the algorithm have (up to sign) similar values.

Within the algorithm the correlation based estimator is used for subsets of varying size. The tuning parameter λ that is used has to be adapted to the number of regressors. For subsets containing only one variable the tuning parameter is λ. For larger subsets we use the penalty

$$\begin{aligned} P_{c,S}(\boldsymbol{\beta}) &= \lambda_{|S|} \sum_{\substack{i<j \\ (i,j)\in S}} \left\{ \frac{(\beta_i - \beta_j)^2}{1-\varrho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1+\varrho_{ij}} \right\} \\ &= 2\lambda_{|S|} \sum_{\substack{i<j \\ (i,j)\in S}} \frac{\beta_i^2 - 2\varrho_{ij} + \beta_j^2}{1-\varrho_{ij}^2} \qquad (7) \end{aligned}$$

where $\lambda_{|S|}$ is a tuning parameter that depends only on the size of S, denoted by $|S|$. In order to have just one tuning parameter the parameter $\lambda_{|S|}$ is chosen as a function of λ. If one considers the case of uncorrelated variables the penalty for all variables given in (2) reduces to $P_c(\boldsymbol{\beta}) = 2\lambda(p-1)\sum_{i=1}^{p}\beta_i^2$ which is the ridge penalty with tuning parameter $2\lambda(p-1)$. Thus $\lambda_{|S|}$ in (7) is chosen by $\lambda_{|S|} = \lambda(|S|-1)$.

Before investigating the performance in high-dimensional settings we demonstrate the grouping effect in a small simulation and consider variable selection for real-life data.

4.1 The grouping effect

For the illustration of the grouping effect we use the idealized example given by Zou & Hastie (2005). With Z_1 and Z_2 being two independent $U(0,20)$ variables the response is generated as $N(Z_1 + 0.1Z_2, 1)$. It is assumed that one observes only

$$\begin{aligned} &\mathbf{x}_1 = Z_1 + \epsilon_1, \quad \mathbf{x}_2 = -Z_1 + \epsilon_2, \quad \mathbf{x}_3 = Z_1 + \epsilon_3, \\ &\mathbf{x}_4 = Z_2 + \epsilon_4, \quad \mathbf{x}_5 = -Z_2 + \epsilon_5, \quad \mathbf{x}_6 = Z_2 + \epsilon_6, \end{aligned}$$

where ϵ_i are independent identically distributed $N(0, 1/16)$. The variables x_1, x_2 and x_3 may be considered as forming one group and x_4, x_5, x_6 as forming a second group. Figure 3 shows the coefficient build-ups for the lasso and BlockBoost for sample size $n = 100$. It is seen that BlockBoost selects the variables x_1, x_2 and x_3 and the corresponding estimates are (up to sign) identical. The strong group consistency of x_1, x_2 and x_3 is distinctly identified. Lasso shows quite different coefficient build-ups selecting as strongly influential the variables x_1 and x_3 and, with rather weak effect, x_2. While coefficient paths for BlockBoost reflect the high correlation of x_1, x_2 and x_3 the path of the lasso are rather irregular. Elastic net behaves quite similar to BlockBoost (compare Zou & Hastie (2005)).

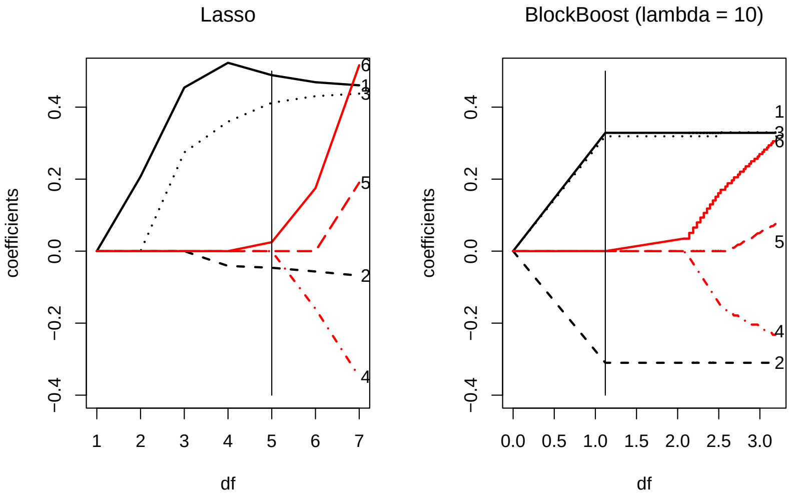


FIGURE 3: *Coefficient build ups for lasso (left) and BlockBoost (right) of the hidden factors example.*

4.2 Application to body fat data

The body fat data set has been used by Penrose, Nelson & Fisher (1985). The study aims at the estimation of the percentage of body fat by various body circumference measurements for 252 men. The thirteen regressors are age (1), weight (lbs) (2), height (inches) (3), neck circumference (4), chest circumference (5), abdomen 2 circumference (6), hip circumference (7), thigh circumference (8), knee circumference (9), ankle circumference (10), biceps (extended) circumference (11), forearm circumference (12), and wrist circumference (13). All circumferences are measured in cm. The percent body fat has been calculated from the equation by Siri (1956) using the body density determined by underwater weighting.

In order to investigate the performances of the alternative approaches the data set has been split 20 times at random into a training set of 151 observations and a test set of 101 observations. Tuning parameters have been chosen by tenfold cross validation. The performance in terms of the median mean squared errors is given in Table 2, the corresponding boxplots are shown in Figure 4. It is seen that correlation based penalization has the best performance in terms of mean squared errors, BlockBoost and elastic net select the same number of variables.

Figure 5 shows the coefficient build-ups for lasso, elastic net, ridge regression, correlation based estimation and BlockBoost based on the full data set. It is seen that the paths for ridge regression and the correlation based estimator are very similar. There is also some similarity between the paths of the elastic net and the lasso. BlockBoost selects 5 variables whereas the elastic net and lasso select 9 and 11 variables, respectively. The reduction to relevant variables is about

Method	median MSE_y	median no. of selected variables
Ridge regression	20.84	13
Lasso	21.80	9.5
Elastic net	24.38	6
CP	20.67	13
BlockBoost	21.70	6

TABLE 2: *Body fat Data - median test mean squared error over 20 random splits for different methods.*

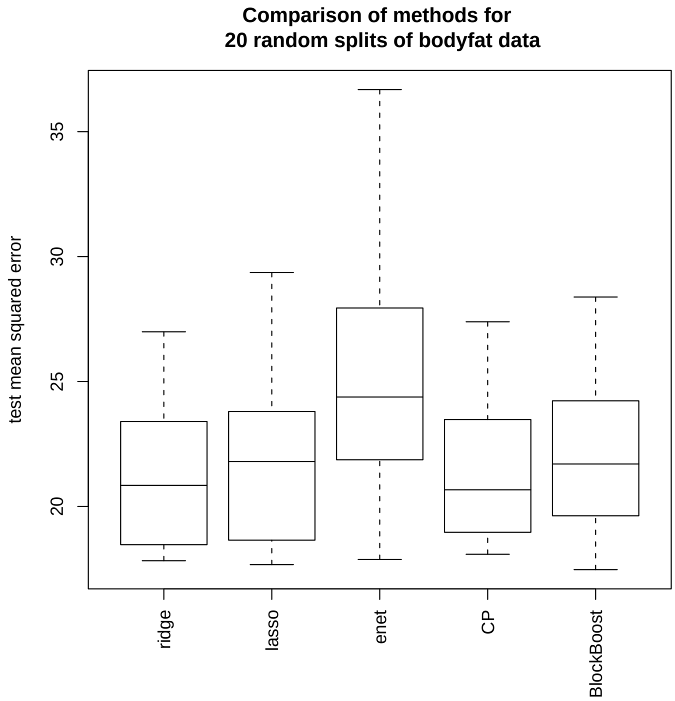


FIGURE 4: *Boxplots of different methods for 20 random splits of body fat data set into a training set of 151 observations and a test set of 101 observations.*

the same for both procedures whereas BlockBoost includes less variables into the final predictor. The estimates given in Table 3 show that strong differences are only found for variables 4, 12 and 13.

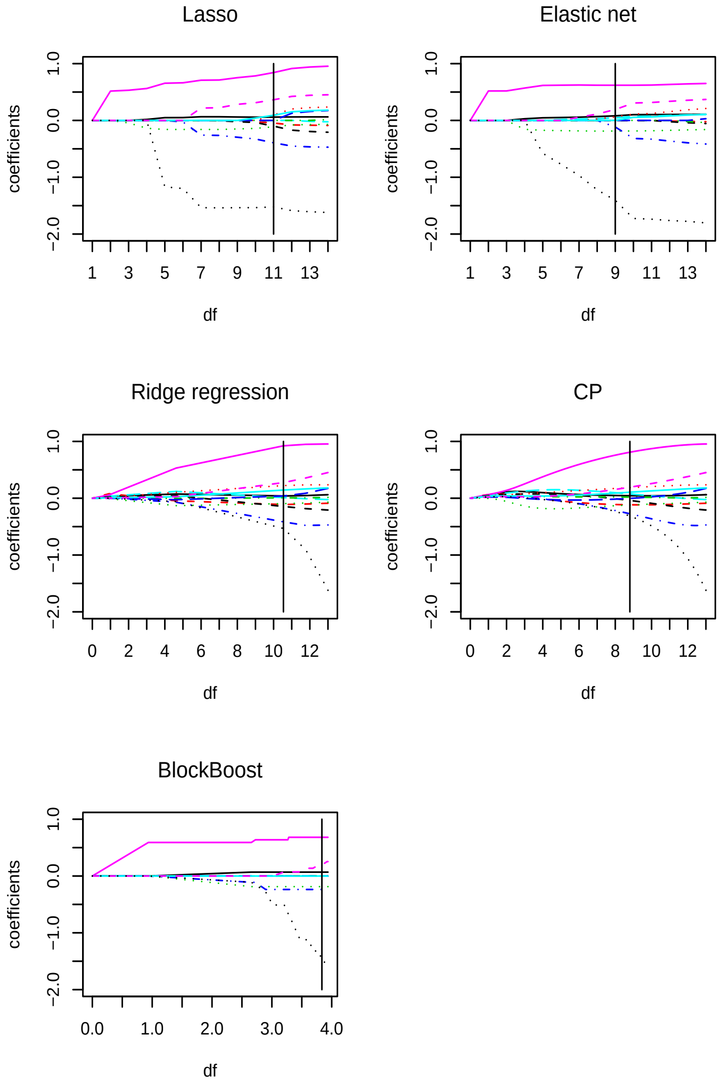


FIGURE 5: *Coefficient build-ups for body fat data based on five estimation methods: lasso (upper left panel), elastic net (upper right panel), ridge regression (mid left panel), correlation based estimation and BlockBoost (lower left panel).*

5 Performance in high-dimensional settings

5.1 Predictive power and estimation of effects

In the following the same notation is used as for the simulations in Section 3. However, the focus is now on high dimensional problems with many predictors. We use the following three high dimensional simulation scenarios.

Variables	Ridge	Lasso	Elastic net	CP	BlockBoost
Tuning	$\lambda = 148.41$	$s = 0.79$	$\lambda = 0.05$	$\lambda = 8.17$	$\lambda = 2000$
parameters:			$s = 0.77$		$m = 56$
1	0.07	0.06	0.09	0.08	0.09
2	-0.03	-0.05	0	-0.03	0
3	-0.16	-0.11	-0.19	-0.18	-0.15
4	-0.43	-0.40	-0.24	-0.28	-0.11
5	0.05	0	0.06	0.1	0
6	0.77	0.86	0.63	0.65	0.69
7	-0.16	-0.11	0	-0.03	0
8	0.19	0.12	0.09	0.13	0
9	0.10	0	0	0.08	0
10	0.02	0.02	0	-0.002	0
11	-0.04	0.10	0.03	-0.05	0
12	0.01	0.37	0.26	-0.04	0
13	-0.39	-1.53	-1.60	-0.242	-1.29

TABLE 3: *Body fat data - tuning parameters and estimated parameters for the whole data set.*

(H1) The parameter vector in the first high-dimensional simulation scenario is given by $\boldsymbol{\beta}^T = (\underbrace{3,\ldots,3}_{5}, \underbrace{0,\ldots,0}_{5}, \underbrace{3,\ldots,3}_{5}, \underbrace{0,\ldots,0}_{5}, \underbrace{3,\ldots,3}_{5}, \underbrace{0,\ldots,0}_{25})$. The correlation $\varrho(x_i, x_j) = \varrho_{ij}$ is given by

$$\varrho_{ij} = \begin{cases} 1 - 0.01 \cdot |i-j|, & i,j \in \{k, k+1, \ldots, k+4\},\ k \in \{1,6,11,16,21\} \\ \epsilon_{ij}, & \text{otherwise,} \end{cases}$$

where ϵ_{ij} are truncated *iid* $N(0, 0.1^2)$. The simulated data has 20|20|40 for training set, independent validation set and test set.

(H2) The parameter vector and the setting is the same as in H1, with weaker correlation specified by

$$\varrho_{ij} = 1 - 0.05 \cdot |i-j|, \quad i,j \in \{k, k+1, \ldots, k+4\},\ k \in \{1,6,11,16,21\}.$$

(H3) The parameter vector is given by

$$\boldsymbol{\beta}^T = (5,4,3,2,1, \underbrace{0,\ldots,0}_{5}, 5,4,3,2,1, \underbrace{0,\ldots,0}_{5}, -5,-4,-3,-2,-1, \underbrace{0,\ldots,0}_{25}),$$

and the correlation $\varrho(x_i, x_j) = \varrho_{ij}$ is given by

$$\varrho_{ij} = \begin{cases} 1 - 0.075 \cdot |i-j|, & i,j \in \{k, k+1, \ldots, k+4\},\ k \in \{1,6,11,16,21\} \\ \epsilon_{ij}, & \text{otherwise,} \end{cases}$$

where ϵ_{ij} are truncated *iid* $N(0, 0.1^2)$. The evaluation setting is 20|20|40.

In the simulations an additional variant of BlockBoost, called BlockBoost(cut), is investigated. It is designed to delete variables that have been selected only once or twice within the iterative selection procedure. More concrete, predictor i is deleted, if $|\beta_i|/\sum_j |\beta_j| < 0.01$. Again we consider the prediction MSE_y and the mean squared error for the estimator of β, MSE_β.

Method	Simulation H1		Simulation H2		Simulation H3	
	median MSE_y	median MSE_β	median MSE_y	median MSE_β	median MSE_y	median MSE_β
Ridge Regression	48.11	27.27	72.77	44.33	63.35	54.06
LASSO	19.62	131.63	59.86	88.06	20.90	125.51
Elastic Net	**14.94**	97.28	**38.55**	51.31	**17.34**	87.84
CP	47.63	26.48	72.99	**44.32**	64.08	53.96
BlockBoost	23.01	**19.36**	67.10	44.99	36.01	38.11
BlockBoost (cut)	23.53	19.67	69.54	45.15	35.02	**38.00**

TABLE 4: *Median test mean squared errors and median MSE_β for the simulated examples (H1), (H2) and (H3) and six methods based on 50 replications.*

The simulation results are given in Table 4 and Figure 6. In all three settings, the elastic net has the best prediction, followed by the lasso. However, if one considers the accuracy of the parameter estimate, the performance of BlockBoost is distinctly superior to the elastic net and the lasso. BlockBoost seems to dominate in terms of MSE_β. One reason is that BlockBoost does somewhat better in identifying relevant variables. This effect is investigated more closely in the next section.

5.2 Identification of relevant variables

While prediction performance is an important criterion for comparison of methods the variables included into the final model are of special interest to practitioners. The final model should be as parsimonious as possible but all relevant variables should be included. The criteria by which the performance of procedures can be judged are the *hits* (i.e. the number of correctly identified influential variables) and the false positives (i.e. the number of non-influential variables dubbed influential). Table 5 and Figure 7 show the mean hits and false positives for the high dimensional simulation settings. Figure 7 is constructed in a way that is similar to ROC curves, but showing points rather than curves. The ideal case performance

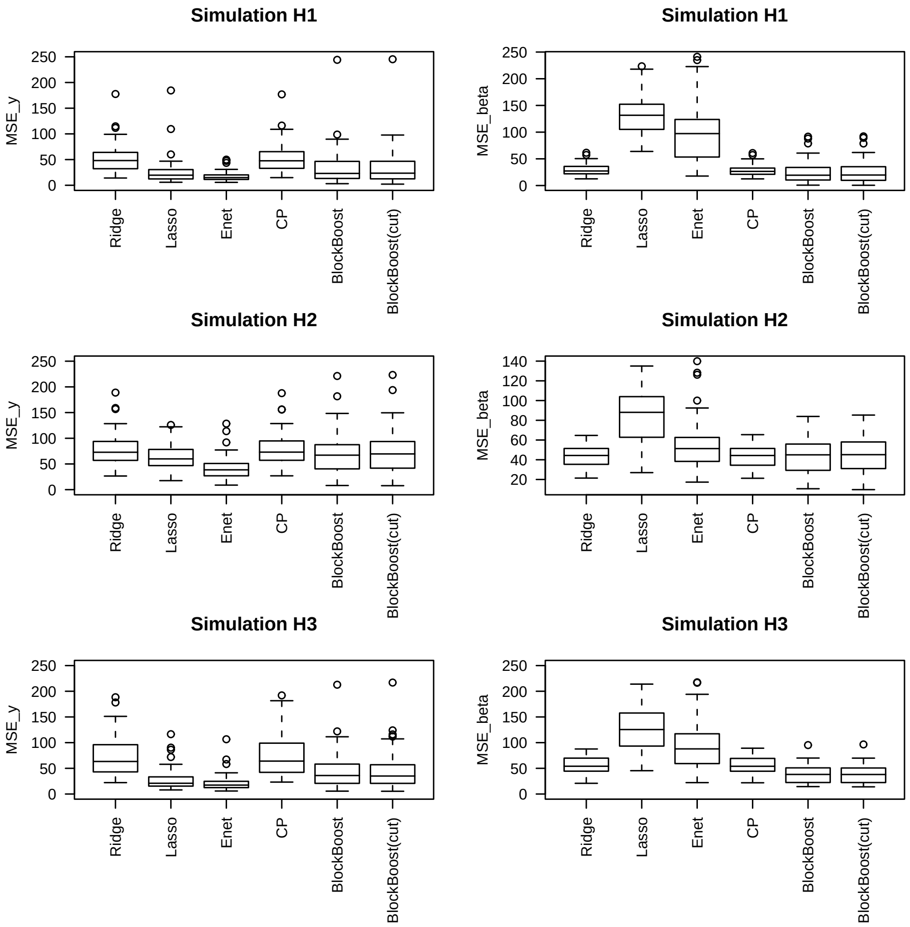


Figure 6: *Boxplots of test mean-squared errors (left column) and MSE_β (right column) for simulations H1 to H3.*

corresponds to the left upper corner. Deviations to the right correspond to an increase in false positives; bad performance in identification of relevant variables corresponds to low values on the abscissae. Due to construction, ridge and correlation based estimator are found in the right upper corner meaning all relevant variables are included but also all irrelevant variables. BlockBoost (as well as BlockBoost(cut)) performs very well in identifying relevant variables. Lasso definitely misses some of the relevant variables but also elastic net has a tendency to miss some. In terms of false positives BlockBoost is comparable to elastic net and lasso while BlockBoost(cut) has the best performance in terms of false positives. The same tendency may be seen from Figure 8 where the performance of the elastic net, the lasso and BlockBoost are shown for the single simulations. (In Figure 8 the points have been jittered in order to show all the simulated data sets.)

Method	Results for the following examples:					
	Example H1		Example H2		Example H3	
	hits	false positives	hits	false positives	hits	false positives
Ridge regression	15	35	15	35	15	35
Lasso	9	6	11	7	10	8
Elastic net	12	6	14	16	12	8
CP	15	35	15	35	15	35
BlockBoost	15	9	15	11	15	9
BlockBoost (cut)	15	4	13	7	14	5

TABLE 5: *Median number of correctly chosen coefficients for examples H1, H2 and H3.*

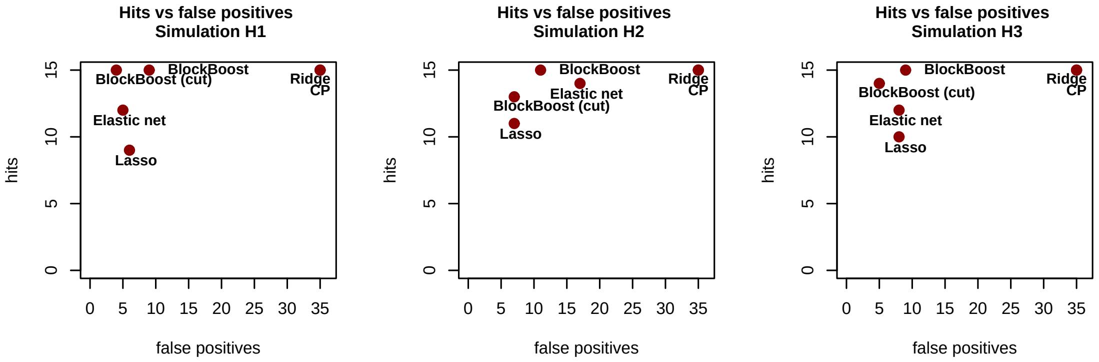


FIGURE 7: *Median hits versus median false positives for simulations H1 to H3.*

For the performance in a real data set we consider the body fat data again. Figure 9 shows the selected variables for BlockBoost and elastic net for the 20 splittings of the data set. It is seen that BlockBoost has less variability in the selection of relevant variables. For example BlockBoost never selects variable 2 whereas it has been selected in four cases by elastic net. Variable 13 has always been selected by BlockBoost but only in 50 percent of the cases by elastic net. Let $h_i, i = 1, \ldots, 13$ denote the number of splits when variable i has been selected. Considering $h_1, \ldots, h_{13}$ as measurements one may compare the standard deviations across the measurements. One obtains 8.55 for BlockBoost and 6.37 for elastic net which shows that BlockBoost is more stable in the sense that it tends to select the same variables across the splits.

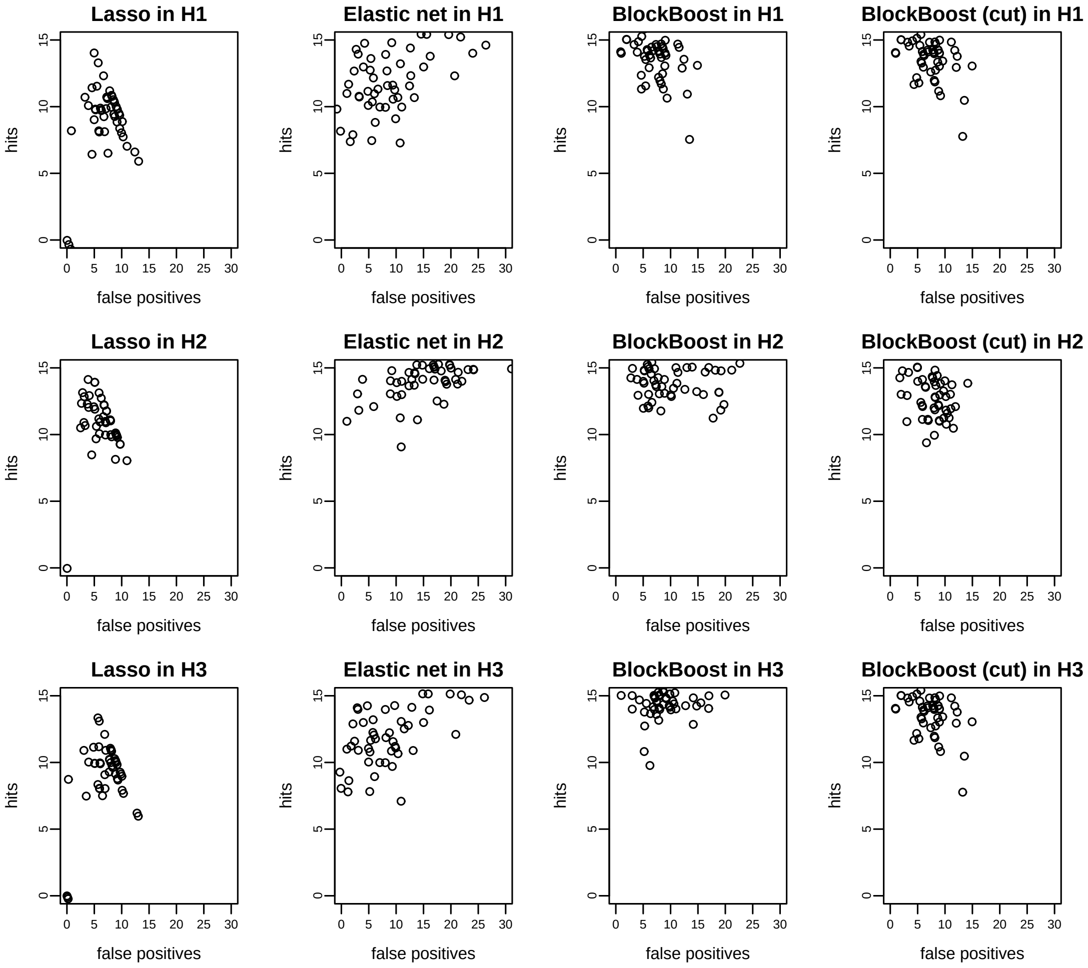


FIGURE 8: *Hits versus false positives for simulations H1 to H3.*

6 Concluding remarks

Two algorithms for the fitting of linear models have been proposed which like the elastic net focus on the grouping effect. It has been demonstrated that, although the elastic net has advantages in predictive power, the correlation based algorithms seem to have superior performance if success is measured by the correct identification of relevant variables. Since in applications, in particular in high dimensional problems, the identification of relevant variables is of crucial importance, the method may be considered as a strong competitor in this field. The method may be extended to generalized linear models by using a penalized likelihood approach. For the correlation based penalty approach the extension is straightforward. Boosted versions may be obtained by modifying LogitBoost

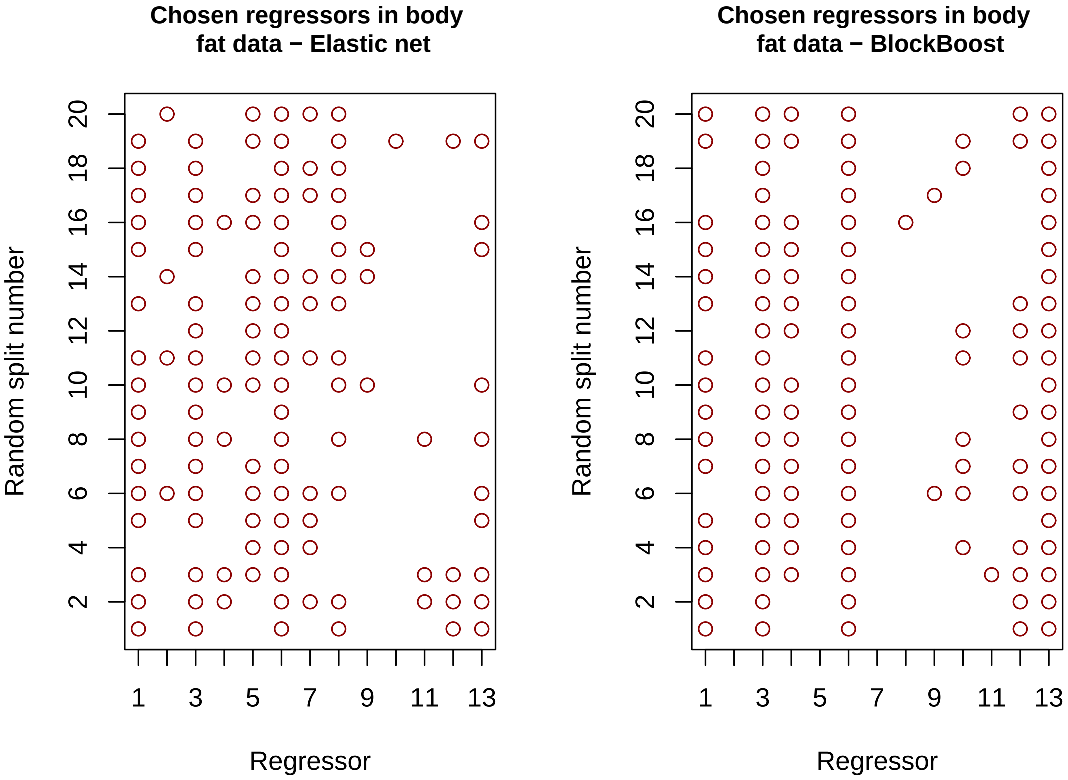


FIGURE 9: *Comparison of chosen regressors between elastic net (left) and BlockBoost (right) for 20 random splits of body fat data.*

(Friedman, Hastie & Tibshirani (1999)).

Appendix

A1: Proposition 1

The penalty P from (2) can be written as

$$P_c = \lambda(\boldsymbol{\beta}^T \mathbf{D^T W_1 D}\boldsymbol{\beta} + \boldsymbol{\beta}^\mathbf{T} \mathbf{A^T W_2 A}\boldsymbol{\beta})$$

where $\mathbf{W}_1 = diag(1/(1-\varrho_{12}), 1/(1-\varrho_{13})\dots)$ is a $(m \times m)$ diagonal matrix, with $m = n(n-1)/2$ denoting the numbers of pairs $(i,j), i \neq j, \mathbf{W}_2 = diag(1/(1+\varrho_{12}), 1/(1+\varrho_{12}), \dots), \mathbf{D}$ specifies the differences,

$$\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & 0 & \dots \\ 1 & 0 & -1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ 0 & 1 & -1 & 0 & \dots \\ 0 & 1 & 0 & -1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and $\mathbf{A}$ is the matrix that specifies the addition of parameters

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots \\ 1 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ 0 & 1 & 1 & 0 & \dots \\ 0 & 1 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The resulting penalty term takes the form $P_c(\boldsymbol{\beta}) = \lambda\boldsymbol{\beta}^T \mathbf{W}\boldsymbol{\beta}$ where $\mathbf{W} = \mathbf{D^T W_1 D} + \mathbf{A^T W_2 A}$. A simpler form of $\mathbf{W}$ is obtained by computing the derivatives. One obtains

$$\frac{\partial P_c(\boldsymbol{\beta})}{\partial \beta_r} = 4\lambda \sum_{i,r} \frac{1}{1-\varrho_{ir}^2}(\beta_r - \varrho_{ir}\beta_i)$$

and

$$\frac{\partial P_c(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s} = \begin{cases} 4\lambda \sum_{i \neq s} \frac{1}{1-\varrho_{is}^2} & \text{if } r = s \\ -4\lambda \frac{\varrho_{rs}}{1-\varrho_{rs}^2} & \text{if } r \neq s. \end{cases}$$

which gives the form (4).

Since a function is strictly convex if the matrix of second derivatives is positive it is enough to show that the quadratic form $P_c(\boldsymbol{\beta})$ takes value zero only for $\boldsymbol{\beta} = \mathbf{0}$.

For $\varrho_{ij}^2 \neq 1, \lambda > 0$, the penalty $P_c(\boldsymbol{\beta})/(2\lambda)$ may be seen as the quadratic Euclidean norm of the expanded vector

$$\mathbf{v} = \left(\frac{\beta_1 - \beta_2}{\sqrt{1-\varrho_{12}}}, \frac{\beta_1 + \beta_2}{\sqrt{1+\varrho_{12}}}, \frac{\beta_1 - \beta_3}{\sqrt{1-\varrho_{13}}}, \frac{\beta_1 + \beta_2}{\sqrt{1+\varrho_{13}}}, \dots \right)$$

Thus, the norm of $\mathbf{v}$ equals zero only if all components are equal to zero. This is only the case if $\beta_i = 0$ for all i. Therefore $P_c(\boldsymbol{\beta}) > 0$ if $\boldsymbol{\beta} \neq 0$. Thus $P_c(\boldsymbol{\beta})$ is strictly convex, and $\hat{\boldsymbol{\beta}}_c$ exists and is unique.

References

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics* **34**, 559–583.

Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association* **98**, 324–339.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–148.

Friedman, J. H., Hastie, T., and Tibshirani, R. (1999). Additive logistic regression: A statistical view of boosting. *Annals of Statistics* **28**, 337–407.

Fu, W. J. (1998). Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). The elements of statistical learning. *Springer-Verlag, New York, USA*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Hurvich, C. M., Simonoff, J. S., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society* **B 60**, 271–293.

Klinger, A. (1998). *Hochdimensionale Generalisierte Lineare Modelle*. Ph. D. thesis, LMU München. Shaker Verlag, Aachen.

Penrose, K. W., Nelson, A. G., and Fisher, A. G. (1985). Generalized body composition prediction equation for men using simple measurement techniques. *Medicine and Science in Sports and Exercise* **17**, 189.

Siri, W. B. (1956). The gross composition of the body. In C. A. Tobias & J. H. Lawrence (Eds.), *Advances in Biological and Medical Physics*, Volume 4, pp. 239–280. Academic Press New York.

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

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