Tutz, Binder:

# Boosting Ridge Regression

Sonderforschungsbereich 386, Paper 418 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

# Boosting Ridge Regression

Gerhard Tutz[1] & Harald Binder[2]

[1] *Ludwig-Maximilians-Universität München, Germany*

[2] *Universität Regensburg, Germany*

July 2005

## Abstract

Ridge regression is a well established method to shrink regression parameters towards zero, thereby securing existence of estimates. The present paper investigates several approaches to combining ridge regression with boosting techniques. In the direct approach the ridge estimator is used to fit iteratively the current residuals yielding an alternative to the usual ridge estimator. In partial boosting only part of the regression parameters are reestimated within one step of the iterative procedure. The technique allows to distinguish between variables that are always included in the analysis and variables that are chosen only if relevant. The resulting procedure selects variables in a similar way as the Lasso, yielding a reduced set of influential variables. The suggested procedures are investigated within the classical framework of continuous response variables as well as in the case of generalized linear models. In a simulation study boosting procedures for different stopping criteria are investigated and the performance in terms of prediction and the identification of relevant variables is compared to several competitors as the Lasso and the more recently proposed elastic net. For the evaluation of the identification of relevant variables pseudo ROC curves are introduced.

Key words: Ridge regression, boosting, Lasso, Pseudo ROC curves

# 1 Introduction

Ridge regression has been introduced by Hoerl and Kennard (1970b) to overcome problems of existence of the ordinary least squares estimator and achieve better prediction. In linear regression with $y = X\beta + \epsilon$ where $y$ is a $n$-vector of centered responses, X an $(n \times p)$-design matrix, $\beta$ a $p$-vector of parameters and $\epsilon$ a vector of iid random errors, the estimator is obtained by minimizing the least squares $(y - X\beta)^T(y - X\beta)$, subject to a constraint $\sum_j |\beta_j|^2 \leq t$. It has the explicit form $\beta_R = (X^T X + \lambda I_p)^{-1} X^T y$ where the tuning parameter $\lambda$ depends on $t$ and $I_p$ denotes the $(p \times p)$ identity matrix. The shrinkage towards zero makes $\beta_R$ a biased estimator. However, since the variance is smaller than for the ordinary least squares estimator ($\lambda = 0$), better estimation is obtained. For details see Seber (1977), Hoerl and Kennard (1970b,a), Frank and Friedman (1993).

Alternative shrinkage estimators have been proposed by modifying the constraint. Frank and Friedman (1993) introduced bridge regression which is based on the constraint $\sum |\beta_j|^\gamma \leq t$ with $\gamma \geq 0$. Tibshirani (1996) proposed the Lasso which results for $\lambda = 1$ and investigated its properties. A comparison between bridge regression and the Lasso as well as a new estimator for the Lasso is found in Fu (1998). Extensions of the ridge estimator to generalized linear models have been considered by Le Cessie and van Houwelingen (1992).

In the present paper boosted versions of the ridge estimator are investigated. Boosting has been originally developed in the machine learning community as a mean to improve classification (e.g. Schapire, 1990). More recently it has been shown that boosting can be seen as the fitting of an additive structure by minimizing specific loss functions (Breiman, 1999; Friedman et al., 2000). Bühlmann and Yu (2003) and Bühlmann (2005) have proposed and investigated boosted estimators in the context of linear regression with the focus on $L2$ loss. In the following linear as well as generalized linear models (GLMs) are considered. In the linear case boosting is based an $L2$ loss whereas in the more general case of GLMs maximum likelihood based boosting is used. In Section 2 we consider simple boosting techniques where the ridge estimator is used iteratively to fit residuals. It is shown that although the type of shrinkage differs for ridge regression and boosted ridge regression, in practice the performance of the two estimators is quite similar. In addition boosted ridge regression inherits from ridge regression that all the predictors are kept in the model and therefore never produces a parsimonious model. In Section 3 we modify the procedure by introducing partial boosting. Partial boosting is a parsimonious modelling strategy like the Lasso (Tibshirani, 1996) and the more recently proposed elastic net (Zou and Hastie, 2005). These approaches are appealing since they produce sparse representations and accurate prediction. Partial boosting means that only a selection of the regression parameters is reestimated in one step. The procedure has several advantages: by estimating only components of the predictor the performance is strongly improved; important parameters are automatically selected and moreover the procedure allows to distinguish between mandatory variables, for example the treatment effect in treatment studies, and optional variables, for example covariates which might be of relevance. The inclusion of mandatory variables right from the beginning yields coefficients paths that are quite different from the paths resulting for the Lasso or the elastic net. An example is given in Figure 6 where two variables ("persdis" and "nosec") are included as mandatory in a binary response example (details are given in Section 6). An advantage of boosted ridge regression is that it can be extended to generalized linear model settings. In Section 4 it is shown that generalized ridge boosting may be constructed as likelihood based boosting including variable selection and customized penalization techniques. The performance of boosted ridge regression for continuous response variable and generalized linear model settings is investigated in Section 5 by simulation techniques. It is demonstrated that componentwise ridge boosting behaves often similar to the Lasso and elastic net with which it shares the property of automatic variable selection. A special treatment is dedicated to the identification of influential variables which is investigated by novel pseudo ROC curves. It is shown that for moderate correlation among covariates boosted ridge estimators perform very well. The paper concludes with a real data example.

## 2 Boosting in ridge regression with continuous response

### 2.1 The estimator

Boosting has been described in a general form as forward stepwise additive modelling (Hastie et al., 2001). Boosted ridge regression as considered here means that the ridge estimator is applied iteratively to the residuals of the previous iteration. For the linear model the algorithm has the simple form:

**Algorithm: BoostR**

Step 1: Initialization

$$\hat{\beta}_{(0)} = (X^T X + \lambda I_p)^{-1} X^T y\,,\ \ \hat{\mu}_{(0)} = X\hat{\beta}_{(0)}$$

Step 2: Iteration

For $m = 1, 2, \dots$ apply ridge regression to the model for residuals

$$y - \hat{\mu}_{(m-1)} = X\beta^R + \varepsilon,$$

yielding solutions

$$\hat{\beta}^R_{(m)} = (X^T X + \lambda I_p)^{-1} X^T (y - \hat{\mu}_{(m-1)}),$$

$$\hat{\mu}^R_{(m)} = X\hat{\beta}^R_{(m)}.$$

The new estimate is obtained by

$$\hat{\mu}_{(m)} = \hat{\mu}_{(m-1)} + \hat{\mu}^R_{(m)}.$$

The procedure may be motivated by stagewise functional gradient descend (Friedman, 2001; Bühlmann and Yu, 2003). Within this framework one wants to minimize, based on data $(y_i, x_i)$, $i = 1, \dots, n$, the expected loss $E\{C(y, \mu(x))\}$ where $C(y, \mu) \geq 0$ is a loss function and $\mu(x)$ represents an approximation to $E(y|x)$. In order to minimize the expected loss an additive expansion of simple learners (fitted functions) $h(x, \hat{\theta})$ is used where $\theta$ represents a parameter. After an initialization step that generates $\hat{\mu}_{(0)}$, in an iterative way the negative gradient vector $u^T = (u_1, \dots, u_n)$, $u_i = -\partial C(y_i, \mu)/\partial\mu|_{\mu=\mu_i^{(m)}}$

is computed and the real learner is fit to the data $(y_i, u_i)$, $i = 1, \ldots, n$. In boosted ridge regression the learner is ridge regression, i.e. $\theta$ is given by $\beta$ and the estimate is $\hat{\beta}_u = (X^T X + \lambda I)^{-1} X u$. If $L2$ loss $C(y, \mu) = (y - \mu)^2/2$ is used, $u_i$ has the simple form of a residual since $-\partial C(y, \mu)/\partial \mu = y - \mu$. Therefore one obtains the BoostR algorithm.

The basic algorithm is completed by specifying a stopping criterion. Since the stopping criterion depends on the properties of the estimates, it will be treated later. In the following properties of the estimator after $m$ iterations are investigated.

## 2.2 Effects of ridge boosting

Simple ridge regression is equivalent to the initialization step with tuning parameter $\lambda$ chosen data-adaptively, for example by cross-validation. Boosted ridge regression is based on an iterative fit where in each step a weak learner is applied, i.e. $\lambda$ is chosen very large. Thus in each step the improvement of the fit is small. The essential tuning parameter is the number of iterations. As will be shown below, the iterative fit yields a tradeoff between bias and variance, which differs from the tradeoff found in simple ridge regression.

It is straightforward to derive that after $m$ iterations of BoostR one has

$$\hat{\mu}_{(m)} = X \hat{\beta}_{(m)}$$

where

$$\hat{\beta}_{(m)} = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}^R_{(j)}$$

shows the evolution of the parameter vector by successive correction of residuals.

It is helpful to have the estimated mean $\hat{\mu}_{(m)}$ and parameter $\hat{\beta}_{(m)}$ after $m$ iterations in closed form. With $B = (X^T X + \lambda I_p)^{-1} X^T$, $S = XB$ one obtains by using Proposition 1 from Bühlmann and Yu (2003)

$$\hat{\mu}_{(m)} = \sum_{j=0}^{m} S(I_n - S)^j y = (I_n - (I_n - S)^{m+1}) y, \tag{1}$$

$$\hat{\beta}_{(m)} = \sum_{j=0}^{m} B(I_n - S)^j y. \tag{2}$$

Thus $\hat{\mu}_{(m)} = H_m y$ with the hat matrix given by

$$\begin{aligned} H_m &= I_n - (I_n - S)^{m+1} \\ &= I_n - (I_n - X(X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}})^{m+1}. \end{aligned}$$

Some insight into the nature of boosted ridge regression may be obtained by using the singular value decomposition of the design matrix. The singular value decomposition of the $(n \times p)$ matrix $X$ with $rank(X) = r \leq p$, is given by $X = UDV^T$, where the $(n \times p)$ matrix $U = (u_1, \ldots, u_p)$ spans the column space of $X$, $U^{\mathrm{T}}U = I_p$, and the $(p \times p)$ matrix $V$ spans the row space of $X$, $V^{\mathrm{T}}V = VV^{\mathrm{T}} = I_p$. $D = diag(d_1, \ldots, d_r, 0, \ldots, 0)$ is a $(p \times p)$ diagonal matrix with entries $d_1 \geq d_2 \geq \cdots \geq d_r \geq 0$ called the singular values of $X$.

One obtains for $\lambda > 0$

$$\begin{aligned} H_m &= I_n - (I_n - UDV^{\mathrm{T}}(VDU^{\mathrm{T}}UDV^{\mathrm{T}} + \lambda I_p)^{-1}VDU^{\mathrm{T}})^{m+1} \\ &= I_n - (I_n - UDV^{\mathrm{T}}(VD^2V^{\mathrm{T}} + \lambda I_p)^{-1}VDU^{\mathrm{T}})^{m+1} \\ &= I_n - (I_n - UD(D^2 + \lambda I_p)^{-1}DU^{\mathrm{T}})^{m+1} \\ &= I_n - (I_n - U\widetilde{D}U^{\mathrm{T}})^{m+1}, \end{aligned}$$

where $\widetilde{D} = diag(\tilde{d}_1^2, \ldots, \tilde{d}_p^2)$, with $\tilde{d}_j^2 = d_j^2/(d_j^2 + \lambda)$, $j = 1, \ldots, r$, $\tilde{d}_j^2 = 0$, $j = r+1, \ldots, p$. Simple derivation shows that

$$H_m = U(I_n - (I_n - \widetilde{D})^{m+1})U^{\mathrm{T}} = \sum_{j=1}^{p} u_j u_j^{\mathrm{T}}(1 - (1 - \tilde{d}_j^2)^{m+1}).$$

It is seen from

$$\hat{\mu}_{(m)} = H_m y = \sum_{j=1}^{p} u_j(1 - (1 - \tilde{d}_j^2)^{m+1})u_j^{\mathrm{T}} y$$

that the coordinates with respect to the orthonormal basis $u_1, \ldots, u_p$, given by $(1 - (1 - \tilde{d}_j^2)^{m+1})u_j^T y$, are shrunken by the factor $(1 - (1 - \tilde{d}_j^2)^{m+1})$. This shrinkage might be compared to shrinkage in common ridge regression. The shrinkage factor in ridge regression with tuning parameter $\gamma$ is given by $d_j^2/(d_j^2 + \gamma)$. If $\lambda \neq 0$, usually no values

$\gamma$, $m$ exist such that $d_j^2/(d_j^2+\gamma) = (1-(1-d_j^2/(d_j^2+\lambda))^{m+1})$ for $j = 1, \ldots, p$. Therefore boosting ridge regression yields different shrinkage than usual ridge regression.

Basic properties of the boosted ridge estimator may be summarized in the following proposition (for proof see appendix).

**Proposition 1.** *(1) The variance after $m$ iterations is given by*

$$cov(\hat{\mu}_{(m)}) = \sigma^2 U(I-(I-\widetilde{D})^{m+1})^2 U^T.$$

*(2) The bias $b = E(\hat{\mu}_{(m)} - \mu)$ has the form*

$$b = U(I-\tilde{D})^{m+1} U\mu.$$

*(3) The corresponding mean squared error is given by*

$$\begin{aligned} MSE(BoostR_\lambda(m)) &= \frac{1}{n}(trace(cov(H_{(m)}y)) + b^T b) \\ &= \frac{1}{n}\sum_{j=1}^{p}\{\sigma^2(1-(1-\tilde{d}_j^2)^{m+1})^2 + c_j(1-\tilde{d}_j^2)^{2m+2})\} \end{aligned}$$

*where $\tilde{d}_j^2 = d_j^2/(d_j^2+\lambda)$, and $c_j = \mu^T u_j u_j^T \mu = \|\mu^T u_j\|$ depends only on the underlying model.*

As consequence, the variance component of the MSE increases exponentially with $(1-(1-\tilde{d}_j^2)^{(m+1)})^2$ whereas the bias component $b^T b$ decreases exponentially with $(1-\tilde{d}_j^2)^{(2m+2)}$. By rearranging terms one obtains the decomposition

$$MSE(BoostR_\lambda(m)) = \frac{1}{n}\sum_{j=1}^{p}\sigma^2 + (1-\tilde{d}_j^2)^{m+1}\{(\sigma^2+c_j)(1-\tilde{d}_j^2)^{m+1} - 2\sigma^2\}.$$

It is seen that for large $\lambda$ $(1-\tilde{d}_j^2$ close to 1) the second term is influential. With increasing iteration number $m$ the positive term $(\sigma^2+c_j)(1-\tilde{d}_j^2)^{m+1}$ decreases while the negative term $-2\sigma^2$ is independent of $m$. Then for appropriately chosen $m$ the decrease induced by $-2\sigma^2$ becomes dominating. The following Proposition states that the obtained estimator yields better MSE than the ML estimate (for proof see appendix).

**Proposition 2.** *In the case where the least squares estimate exists ($r \leq p$) one obtains*

*(a) For $m \to \infty$ the MSE of boosted ridge regression converges to the MSE of the least squares estimator $r\sigma^2/n$*

*(b) For appropriate choice of the iteration number boosted ridge regression yields smaller values in MSE than the least squares estimates*

Thus if the least squares estimator exists boosted Ridge Regression may be seen as an alternative (although not effective) way of computing the least squares estimate ($m \to \infty$).

For simple ridge regression the trade-off between variance and bias has a different form. For simple ridge regression with smoothing parameter $\gamma$ one obtains

$$MSE(Ridge_\gamma) = \frac{1}{n}\sum_{j=1}^{p}\{\sigma^2(\tilde{d}^2_{j,\gamma})^2 + c_j(1-\tilde{d}^2_{j,\gamma})^2\}$$

where $\tilde{d}^2_{j,\gamma} = d_j^2/(d_j^2+\gamma)$. While the trade-off in simple ridge regression is determined by the weights $(\tilde{d}^2_{j,\gamma})^2$ and $(1-\tilde{d}^2_{j,\gamma})^2$ on $\sigma^2$ and $c_j$, respectively, in boosted ridge regression the corresponding weights are $(1-(1-\tilde{d}_j^2)^{m+1})^2$ and $(1-\tilde{d}_j^2)^{2m+2}$ which crucially depend on the iteration number.

Figure 1 shows the (empirical) bias and variance for example data generated from model (3) which will be introduced in detail in Section 5. The minimum MSE obtainable is slightly smaller for boosted ridge regression, probably resulting from a faster drop-off of the bias component. We found this pattern prevailing in other simulated data examples, but the difference between boosted ridge regression and simple ridge regression usually is not large. Overall the fits obtainable from ridge regression and boosted ridge regression with the penalty parameter and the number of steps chosen appropriately are rather similar.

## 3 Partial and componentwise boosting

Simple boosted regression as considered in Section 2 yields estimates in cases where the ordinary least-squares estimator does not exist. However, as will be shown, in particu-

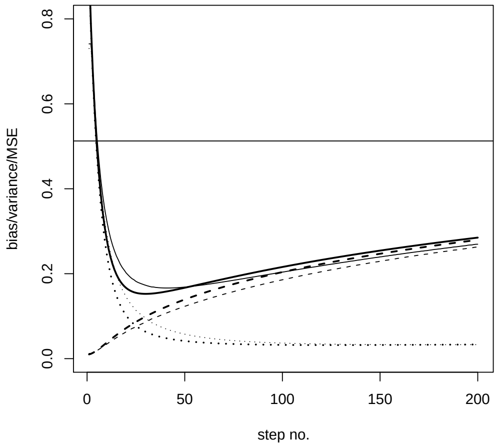


Figure 1: Empirical bias and variance of ridge regression (thin lines) and boosted ridge regression (thick lines) (MSE: solid lines; bias: dotted lines; variance: broken lines) for example data generated from (3) (see Section 5) with $n = 100$, $p = 50$, $\rho_b = 0.7$ and signal-to-noise ratio 1. The horizontal scale of the ridge regression curves is adjusted to have approximately the same degrees of freedom as boosted ridge regression in each step.

lar for high dimensional predictor space the performance can be strongly improved by updating only one component of $\beta$ in one iteration. Bühlmann and Yu (2003) refer to the method as componentwise boosting and propose to choose the component to be updated with reference to the resulting improvement of the fit. This powerful procedure implicitly selects variables to be included in the predictor. What comes as an advantage may turn into a drawback when covariates which are important to the practioner are not included. For example in case-control studies the variables may often be grouped into variables that have to be evaluated, including the treatment, and variables for which inclusion is optional, depending on their effect. Moreover, componentwise boosting does not distinguish between continuous and categorical predictors. In our studies continuous predictors have been preferred in the selection process, probably since continuous predictors contain more information than binary variables.

Therefore in the following *partial boosting* of ridge estimators is proposed where partial boosting means that in one iteration selected components of the parameter vector $\beta_{(m)}^{\mathrm{T}} = (\beta_{(m),1}, \ldots, \beta_{(m),p})$ are re-estimated. The selection is determined by a specific structuring of the parameters (variables). Let the parameter indices $V = \{1, \ldots, p\}$ be partitioned into disjunct sets by $V = V_c \cup V_{o1} \cup \ldots \cup V_{oq}$ where $V_c$ stands for the (compulsory) parameters (variables) that have to be included in the analysis, and $V_{o1}, \ldots, V_{oq}$ represent blocks of parameters that are optional. A block $V_{or}$ may refer to all the parameters which refer to a multicategorical variable, such that not only parameters but variables are evaluated. Candidates in the refitting process are all combinations $V_c \cup V_{or}, r = 1, \ldots, q$, representing combinations of necessary and optional variables. Componentwise boosting, as considered by Bühlmann and Yu (2003) is the special case where in each iteration one component from $\beta_{(m)}$, say $\beta_{(m),j}$, is reestimated, meaning that the structuring is specified by $V_c = \emptyset, V_{oj} = \{j\}$.

Let now $V_m = (m_1, \ldots, m_l)$ denote the indices of parameters to be considered for refitting in the $m$th step. One obtains the actual design matrix from the full design matrix $X = (x_{\cdot 1}, \ldots, x_{\cdot p})$ by selecting the corresponding columns, obtaining the design matrix $X_{V_m} = (x_{\cdot m_1}, \ldots, x_{\cdot m_l})$. Then in iteration $m$ ridge regression is applied to the reduced model

$$y_i - \hat{\mu}_{(m-1),i} = (x_{i,m_1}, \ldots, x_{i,m_l})^T \beta_{V_m}^R + \varepsilon_i$$

yielding solutions

$$\begin{aligned} \hat{\beta}_{V_m}^R &= (\hat{\beta}_{V_m,m_1}^R, \ldots, \hat{\beta}_{V_m,m_l}^R) \\ &= (X_{V_m}^{\mathrm{T}} X_{V_m} + \lambda I_p)^{-1} X_{V_m}^{\mathrm{T}} (y - \hat{\mu}_{(m-1)}). \end{aligned}$$

The total parameter update is obtained from components

$$\hat{\beta}_{(V_m),j}^R = \begin{cases} \hat{\beta}_{V_m,j}^R & j \in V_m \\ 0 & j \notin V_m \ , \end{cases}$$

yielding $\hat{\beta}_{(V_m)}^R = (\hat{\beta}_{(V_m),1}^R, \ldots, \hat{\beta}_{(V_m),p}^R)^T$. The corresponding update of the mean is given by $\hat{\mu}_{(m)} = \hat{\mu}_{(m-1)} + X_{V_m} \hat{\beta}_{V_m}^R = X(\hat{\beta}_{(m-1)} + \hat{\beta}_{(V_m)}^R)$, and the new parameter vector is $\hat{\beta}_{(m)} = \hat{\beta}_{(m-1)} + \hat{\beta}_{(V_m)}^R$.

In the general case where in the $m$th step several candidate sets $V_m^{(j)} = (m_1^{(j)}, \ldots, m_l^{(j)})$, $j = 1, \ldots, l$, are evaluated, an additional selection step is needed. In summary the $m$th iteration step has the following form:

**Iteration ($m$th step): PartBoostR**

(a) Compute for $j = 1, \ldots, s$, the parameter updates $\hat{\beta}_{(V_m^{(j)})}$, and the corresponding means

$$\hat{\mu}_{(m)}^{(j)} = \hat{\mu}_{(m-1)} + X_{V_m^{(j)}} \hat{\beta}_{V_m^{(j)}}.$$

(b) Determine which $\hat{\mu}_{(m)}^{(j)}$, $j = 1, \ldots, l$ improves the fit maximally.

With $V_m$ denoting the subset which is selected in the $m$th step one obtains the updates $\hat{\beta}_{(m)} = \hat{\beta}_{(m-1)} + \hat{\beta}_{(V_m)}^R$, $\hat{\mu}_{(m)} = X\hat{\beta}_{(m)}$. With $S_0 = X(X^TX + \lambda I_p)^{-1}X^T$, $S_m = X_{V_m}(X_{V_m}^T X_{V_m} + \lambda I_p)^{-1} X_{V_m}^T$, $m = 1, 2, \ldots$ the iterations are represented as

$$\hat{\mu}_{(m)} = \hat{\mu}_{(m-1)} + S_m(y - \hat{\mu}_{(m-1)})$$

and therefore

$$\hat{\mu}_{(m)} = H_m y$$

where $H_m = \sum_{j=0}^{m} S_j \prod_{i=1}^{j-1} (I - S_i)$.

If $S_j = S$, $H_m$ simplifies to the form of simple ridge boosting. One obtains $cov(\hat{\mu}_{(m)}) = \sigma^2 H_m^2$, $bias = (H_m - I)\mu$, and the MSE has the form

$$MSE = \frac{1}{n}(trace(\sigma^2 H_m^2) + \mu^T (H_m - I)^2 \mu).$$

While mean squared error (MSE), bias and variance are similar for boosted ridge regression and simple ridge regression (Figure 1), differences can be expected when these approaches are compared to componentwise boosted ridge regression. Figure 2 shows MSE (solid lines), bias (dottes lines) and variance (broken lines) for the same data underlying Figure 1, this time for boosted ridge regression (thick lines) and the componentwise approach (thin lines). Penalties for both procedures have been chosen such that the bias curves are close in the initial steps. The fit of a linear model that

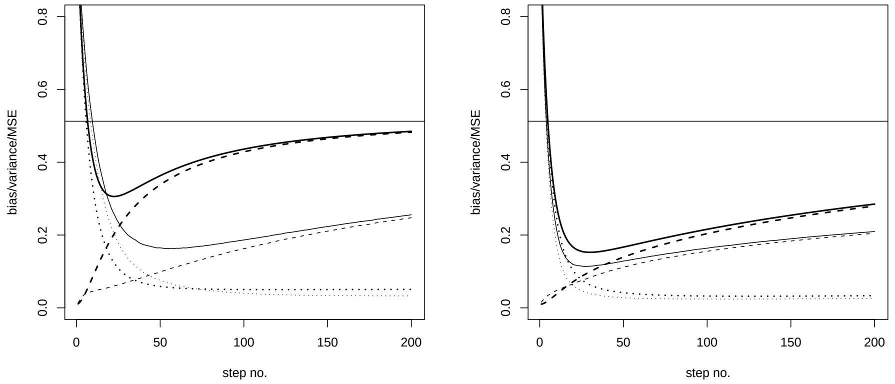


Figure 2: Empirical bias and variance of boosted ridge regression (thick lines) and componentwise boosted ridge regression (thin lines) (MSE: solid lines; bias: dotted lines; variance: broken lines) for example data generated from (3) with $n = 100$, $p = 50$ and uncorrelated ($\rho_b = 0$) (left panel) and correlated ($\rho_b = 0.7$) (right panel) covariates.

incorporates all variables is indicated by a horizontal line. In the left panel of Figure 2 there are five covariates (of 50 covariates in total) with true parameters unequal to zero and the correlation between all covariates is zero. It is seen that for data with a small number of uncorrelated informative covariates the componentwise approach results in a much smaller minimum MSE, probably due to a much slower increase of variance. When correlation among all covariates (i.e. also between the covariates with non-zero parameters and the covariates with true parameters equal to zero) increases, the performance of boosted ridge regression comes closer to the componentwise approach (right panel of Figure 2). This may be due to the coefficient build-up scheme of BoostR (illustrated in Figure 3) that assigns non-zero parameters to all covariates.

The idea of a stepwise approach to regression where in each step just one predictor is updated is also found in stagewise regression (see e.g. Efron et al., 2004). The main difference is the selection criterion and the update: The selection criterion in stagewise regression is the correlation between each variable and the current residual and the update is of fixed size $\epsilon$, whereas in componentwise ridge regression a penalized model with one predictor is fitted for each variable and any model selection criterion may be used. Since stagewise regression is closely related to the Lasso and Least Angle Regression (Efron et al., 2004), it can be expected that componentwise boosted ridge regression is also similar to the latter procedures.

### 3.1 Example: Prostate Cancer

We applied componentwise boosted ridge regression and boosted ridge regression to the prostate cancer data used by Tibshirani (1996) for illustration of the Lasso. The data with $n = 97$ observations come from a study by Stamney et al. (1989) that examined the correlation between the (log-)level of prostate specific antigen and eight clinical measures (standardized before model fit): log(cancer volume) (lcavol), log(prostate weight) (lweight), age, log(benign prostatic hyperplasia amount) (lbph), seminal vesical invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason scores 4 or 5 (pgg45).

Figure 3 shows the coefficient build-up in the course of the boosting steps for two scalings. In the top panels the values on the abscissa indicate the $L_2$-norm of parameter vector relative to the least-squares estimate, whereas in the bottom panels the degrees of freedom are given. A dot on the line corresponding to a coefficient indicates an estimate; thus each dot represents one step in the algorithm. While for boosted ridge regression (left panels) each coefficient seems to increase by a small amount in each step, in the componentwise approach (center panels) only to a specific set of variables non-zero coefficients are assigned within one step, the other coefficients remain zero. Therefore the coefficient build-up of the componentwise approach looks much more similar to that of the Lasso procedure (right panels) (fitted by the LARS procedure, see Efron et al., 2004). One important difference to the Lasso that can be seen in this example is that for BoostPartR the size of the coefficient updates varies over boosting steps in a specific pattern, with larger changes in the early steps and only small adjustments in late steps.

Arrows indicate 10-fold cross-validation based on mean squared error which has been repeated over 10 distinct splittings to show variability. It is seen that cross-validation

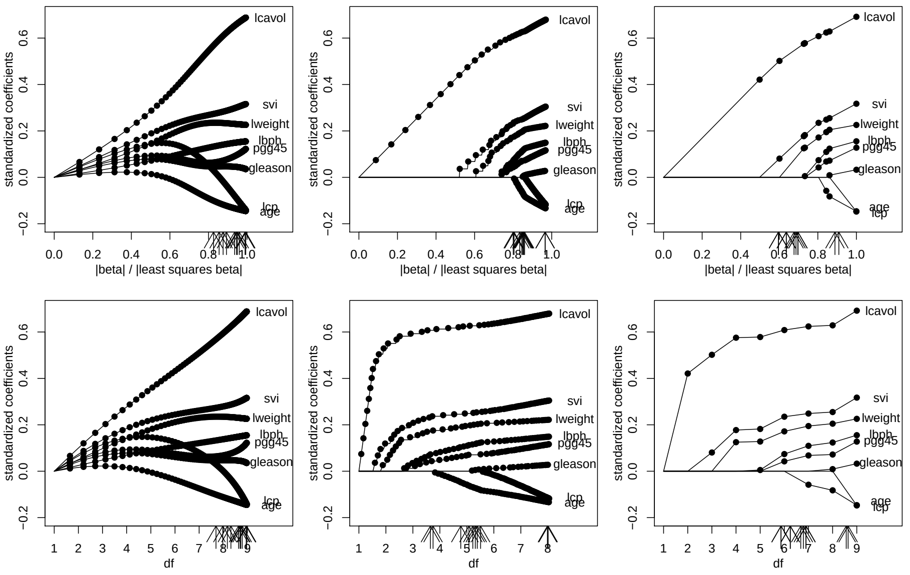


Figure 3: Coefficient build-up of BoostR (left panels), PartBoostR (center panels) and the Lasso (right panels) for the prostate cancer data plotted against standardized $L_2$-norm of the parameter vector (top panels) and degrees of freedom (bottom panels). Arrows indicate the model chosen by 10-fold cross-validation (repeated for 10 times).

selects a model of rather high complexity for boosted ridge regression. When using the componentwise approach or the Lasso, more parsimonious models are selected. Partial boosting selects even more parsimonious models than the Lasso in terms of degrees of freedom.

## 4 Ridge boosting in generalized linear models

In generalized linear models the dominating estimation approach is maximum likelihood which corresponds to the use of $L2$ loss in the special case of normally distributed responses. Ridge regression in generalized linear models is therefore based an penalized maximum likelihood. Univariate generalized linear models are given by

$$\mu_i = h(\beta_0 + x_i^T \tilde{\beta}) = h(z_i^T \beta) = h(\eta_i)$$

where $\mu_i = E(y_i|x_i)$, $h$ is a given (strictly monotone) response function, and the predictor $\eta_i = z_i^T \beta$ has linear form. Moreover, it is assumed that $y_i|x_i$ is from the simple exponential family, including for example binary responses, Poisson distributed responses and gamma distributed responses.

The basic concept in ridge regression is to maximize the *penalized* log-*likelihood*

$$\begin{aligned} l_p(\beta) &= \sum_{i=1}^{n} l_i - \frac{\lambda}{2} \sum_{i=1}^{p} \beta_i^2 \\ &= \sum_{i=1}^{n} l_i - \frac{\lambda}{2} \sum_{i=1}^{p} \beta^T P \beta \end{aligned}$$

where $l_i$ is the likelihood contribution of the ith observation and $P$ is a block diagonal matrix with the $(1 \times 1)$ block 0 and the $(p \times p)$ block given by the identy matrix $I_p$. The corresponding *penalized score function* is given by

$$s_p(\beta) = \sum_{i=1}^{n} z_i \frac{\partial h(\eta_i)/\partial \eta}{\sigma_i^2}(y_i - \mu_i) - \lambda P \beta.$$

where $\sigma_i^2 = var(y_i)$. A simple way of computing the estimator is iterative Fisher scoring

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + F_p(\hat{\beta}^{(k)})^{-1} s_p(\hat{\beta}^{(k)}),$$

where $F_p(\beta) = E(-\partial l/\partial \beta \partial \beta^T) = F(\beta) + \lambda P$, with $F(\beta)$ denoting the usual Fisher matrix given by $F(\beta) = X^T W(\eta) X, X = (x_{\cdot 0}, x_{\cdot 1}, \ldots, x_{\cdot p})$, $x_{\cdot 0}^T = (1, \ldots, 1)$, $W(\eta) = D(\eta)\Sigma(\eta)^{-1}D(\eta)$, $\Sigma(\eta) = (\sigma_1^2, \ldots, \sigma_n^2)$, $D(\eta) = diag(\partial h(\eta_1)/\partial \eta, \ldots, \partial h(\eta_n)/\partial \eta)$.

The proposed boosting procedure is likelihood based ridge boosting based on one step of Fisher scoring (for likelihood based boosting see also Tutz and Binder, 2004). The parameter/variable set to be refitted in one iteration now includes the intercept. Let again $V_m^{(j)}$ denote subsets of indices of parameters to be considered for refitting in the $m$th step. Then $V_m^{(j)} = \{m_1, \ldots, m_l\} \subset \{0, 1, \ldots, p\}$, where 0 denotes the intercept.

The general algorithm, including a selection step, is given in the following. In order to keep the presentation simple the matrix form $\mu = h(\eta)$ is used where $\mu$ and $\eta$ denote vectors in which observations are collected.

**Algorithm: GenBoostR**

Step 1: Initialization

Fit model $\mu_i = h(\beta_0)$ by iterative Fisher scoring obtaining $\hat{\beta}_{(0)} = (\hat{\beta}_0, 0, \ldots, 0)$, $\hat{\eta}_{(0)} = X\hat{\beta}_{(0)}$

Step 2: Iteration

For $m = 1, 2, \ldots$

(a) Estimation

Estimation for candidate sets $V_m^{(j)}$ corresponds to fitting of the model

$$\mu = h(\hat{\eta}_{(m-1)} + X_{V_m^{(j)}} \beta^R_{V_m^{(j)}})$$

where $\hat{\eta}_{(m-1)} = X\hat{\beta}_{(m-1)}$ and $X^T_{V_m^{(j)}} = (x_{im_1^{(j)}}, \ldots, x_{im_l^{(j)}})$ contains only components from $V_m^{(j)}$.

One step of Fisher scoring is given by

$$\hat{\beta}^R_{V_m^{(j)}} = F^{-1}_{p,V_m^{(j)}} s_{p,V_m^{(j)}}$$

where $s_{p,V_m^{(j)}} = X_{V_m^{(j)}} W(\hat{\eta}_{(m-1)}) D^{-1}(\eta_{(m-1)})(y - \mu_{(m-1)})$ (without $-\lambda P\beta = 0$), $F_{p,V_m^{(j)}} = X^T_{V_m^{(j)}} W(\hat{\eta}_{(m-1)}) X_{V_m^{(j)}}$, with $\hat{\eta}^T_{(m-1)} = (\hat{\eta}_{1,(m-1)}, \ldots, \hat{\eta}_{n,(m-1)})$, $y^T = (y_1, \ldots, y_n)$, $\mu^T_{(m-1)} = (\mu_{1,(m-1)}, \ldots, \mu_{n,(m-1)})$.

(b) Selection

For candidate sets $V_m^{(j)}$, $j = 1, \ldots, s$, the set $V_m$ is selected that improves the fit maximally.

(c) Update

One sets

$$\hat{\beta}^R_{(m)} = \begin{cases} \hat{\beta}^R_{V_m,j} & j \in V_m \\ 0 & j \notin V_m, \end{cases}$$

$\hat{\beta}_{(m)} = \hat{\beta}_{(m+1)} + \hat{\beta}^R_{(m)}$,

$\hat{\eta}_{(m)} = X\hat{\beta}_{(m)} = X\hat{\beta}_{(m+1)} + X\hat{\beta}^R_{(m)}$,

$\hat{\mu}_{(m)} = h(X\hat{\beta}_{(m)})$ where $h$ is applied componentwise.

Within a generalized linear model framework it is natural to use in the selection step (b) the improvement in deviance. With $Dev(\hat{\eta})$ denoting the deviance given predictor values $\hat{\eta}$ one selects in the $m$th step $V_m$ such that $Dev(\eta^{(j)})$ is minimal, where $Dev(\eta^{(j)})$ uses predictor value $\hat{\eta}^{(j)} = \hat{\eta}_{m-1} + X_{V_m^{(j)}} \hat{\beta}^R_{V_m^{(j)}}$.

Also stopping criteria should be based on the deviance. One option is deviance based cross-validation, an alternative choice is the AIC criterion

$$AIC = Dev(\hat{\eta}_{(m)}) + 2df_m$$

or the Bayesian information criterion

$$BIC = Dev(\hat{\eta}_{(m)}) + \log n \cdot df_m$$

where $df_m$ represents the effective degrees of freedom which are given by the trace of the hat matrix (see Hastie and Tibshirani, 1990). In the case of generalized linear models the hat matrix is not as straightforward as in the simple linear model. The following proposition gives an approximate hat matrix (for proof see Appendix).

**Proposition 3.** *An approximate hat matrix for which $\hat{\mu}_{(m)} \approx H_m y$ is given by*

$$H_m = \sum_{j=0}^{m} M_j \prod_{i=0}^{j-1} (I - M_0)$$

*where $M_m = \Sigma_m^{1/2} W_m^{1/2} X_{V_m} (X_{V_m}^T W_m X_{V_m} + \lambda I)^{-1} X_{V_m} W_m^{1/2} \Sigma_m^{-1/2}$, $W_m = W(\hat{\eta}_{(m-1)})$, $D_m = D(\hat{\eta}_{(m-1)})$*

The approximate hat matrix yields the AIC criterion $AIC = Dev(\hat{\eta}_{(m)}) + 2tr(H_m)$.

# 5 Empirical comparison

For the investigation of the properties of boosted ridge regression we use simulated data. This allows us the modify systematically the structure of the data fed into the algorithms and to observe how the results change.

We generate a covariate matrix $X$ by drawing $n$ observations from a $p$-dimensional multivariate normal distribution with variance 1 and correlation between two covariates (i.e. columns of $X$) $x_j$ and $x_k$ being $\rho_b^{|j-k|}$. The true predictor $\eta$ (and thereby the corresponding expected value of the response $E(y|x) = \mu = h(\eta)$, where $h$ is the identity for a continuous response and $h(\eta) = \exp(\eta)/(1+\exp(\eta))$ for binary responses) is formed by

$$\eta = X\beta_{true} \tag{3}$$

where the true parameter vector $\beta_{true} = c_{stn} \cdot (\beta_1, \ldots, \beta_p)^T$ is determined by

$$\beta_j \sim N(5,1) \text{ for } j \in V_{info}, \quad \beta_j = 0 \text{ otherwise}$$

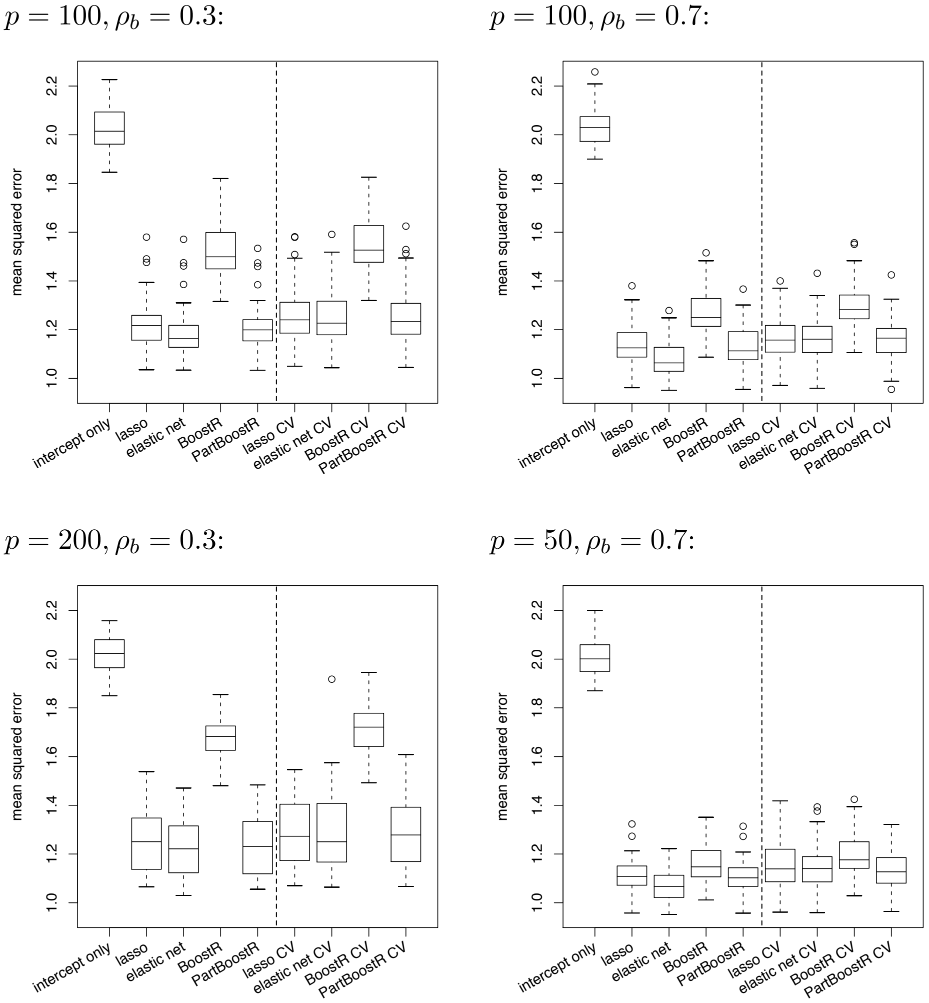
with $V_{info} \subset \{1, \ldots, 10\}$ being the set (of size 5) of the randomly drawn indices of the informative covariates. The constant $c_{stn}$ is chosen such that the signal-to-noise ratio for the final response $y$, drawn from a normal distribution $N(\mu, 1)$ or a binomial distribution $B(\mu, 1)$,is equal to 1. The signal-to-noise-ratio is determined by

$$\text{signal-to-noise ratio} = \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}{\sum_{i=1}^{n} Var(y_i)}$$

where $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i$. For the examples presented here we used fixed sample size $n = 100$ for the training data and a new sample of size $n_{new} = 1000$ for the evaluation of prediction performance.

The following comparisons of performance and identification of influential covariates of (Gen)BoostR and (Gen)PartBoostR with other procedures have been done within the statistical environment R 2.1.0 (R Development Core Team, 2004). We used intercept-only (generalized) linear models, the (generalized) Lasso (package "lasso2" 1.2-0 — LARS as used in Section 3.1 is only available for continuous response data) and the "elastic net" procedure (Zou and Hastie, 2005) (package "elastic net" 1.02) for comparison. We evaluate performance for optimal values of the tuning parameters as well as for parameters chosen by tenfold cross-validation.

For the Lasso only one parameter (the upper bound on the $L_1$-norm of the parameter vector) has to be chosen. We use a line search, which has to be augmented in some examples because for certain values no solution exists. Zou and Hastie (2005) also note this as a downside of the classical Lasso. The elastic net procedure has two tuning


$p = 100, \rho_b = 0.3$:

$p = 100, \rho_b = 0.7$:

$p = 200, \rho_b = 0.3$:

$p = 50, \rho_b = 0.7$:

Figure 4: Mean squared error for continuous response data with varying number of predictors $p$ and correlation $\rho_b$ for the linear model including only an intercept term, elastic net, the Lasso BoostR and PartBoostR with tuning parameters selected for optimal performance or by cross-validation (CV).

parameters, the number of steps $k$ and the penalty $\lambda$. We chose both by evaluating a grid of steps $1, \ldots, min(p, n)$ and penalties $(0, 0.01, 0.1, 1, 10, 100)$ (as suggested by Zou and Hastie, 2005). For each of BoostR and PartBoostR we used a fixed penalty parameter $\lambda$. This parameter has been chosen in advance such that the number of steps chosen by cross-validation typically is in between 50 and 200.

## 5.1 Metric response

### 5.1.1 Prediction performance

The boxplots in Figure 4 show the mean squared error for continuous response data generated from (3) with a varying number of predictors and amount of correlation between the covariates for 50 repetitions per example. In Figure 4 elastic net, the Lasso, boosted ridge regression and componentwise boosted ridge regression (with an intercept-only linear model as a baseline) are compared for the optimal choice of tuning parameters (in the left part of the panels) as well for the cross-validation based estimates (in the right part).

When comparing the optimal performance of boosted ridge regression with the componentwise approach, it is seen that the latter distinctly dominates the former for a large number of predictors (with only few of them being informative) (bottom left vs. top left panel) and/or small correlations (top left vs. top right panel). A similar difference in performance is also found when comparing boosted ridge regression to the Lasso and to the elastic net procedure. This highlights the similarity of the componentwise approach to the Lasso-type procedures (as illustrated in Section 3.1). When the close connection of BoostR and simple ridge regression is taken into consideration this replicates the results of Tibshirani (1996) who also found that in sparse scenarios the Lasso performs well and ridge regression performs poor. Consistently the performance difference for a smaller number of covariates and high correlation (bottom right panel) is less pronounced.

For optimal parameters no large differences between the performance of the componentwise approach, the Lasso and of elastic net are seen. Elastic net seems to have a slightly better performance for examples with higher correlation among the predictors, but this is to be expected because elastic net was specifically developed for such high-correlation scenarios (see Zou and Hastie, 2005). The decrease in performance incurred by selecting the number of steps/Lasso constraint by cross-validation instead of using the optimal values is similar for componentwise boosted ridge regression and the Lasso. For the elastic net procedure the performance decrease is larger to such an extent that the performance benefit over the former procedures (with optimal parameters) is lost. This might result from the very small range of the number of steps where good performance can be achieved (due to the small overall number of steps used by elastic net). In contrast boosting procedures change very slowly in the course of the iterative process, which makes selection of appropriate shrinkage more stable (compare Figure 3).

### 5.1.2 Identification of influential variables

While prediction performance is an important criterion for comparison of algorithms the variables included into the final model are also of interest. The final model should be as parsimonious as possible, but all relevant variables should be included. For example one of the objectives for the development of the elastic net procedure was to retain all important variables for the final model even when they are highly correlated (while the Lasso includes only some of a group of correlated variables; see Zou and Hastie, 2005).

The criteria by which the performance of a procedures in the identification of influential variables can be judged are the *hit rate* (i.e. the proportion of correctly identified influential variables)

$$\text{hit rate} = \frac{1}{\sum_{j=0}^{p} I(\beta_{true,j} \neq 0)} \sum_{j=1}^{p} I(\beta_{true,j} \neq 0) \cdot I(\hat{\beta}_j \neq 0)$$

and the *false alarm rate* (i.e. the proportion of non-influential variables dubbed influential)

$$\text{false alarm rate} = \frac{1}{\sum_{j=0}^{p} I(\beta_{true,j} = 0)} \sum_{j=1}^{p} I(\beta_{true,j} = 0) \cdot I(\hat{\beta}_j \neq 0)$$

where $\beta_{true,j}, j = 1, \dots, p$ are the elements of the true parameter vector, $\hat{\beta}_j$ are the corresponding estimates used by the final model and $I$(expression) is an indicator function that takes the value 1 if “expression” is true and 0 otherwise.

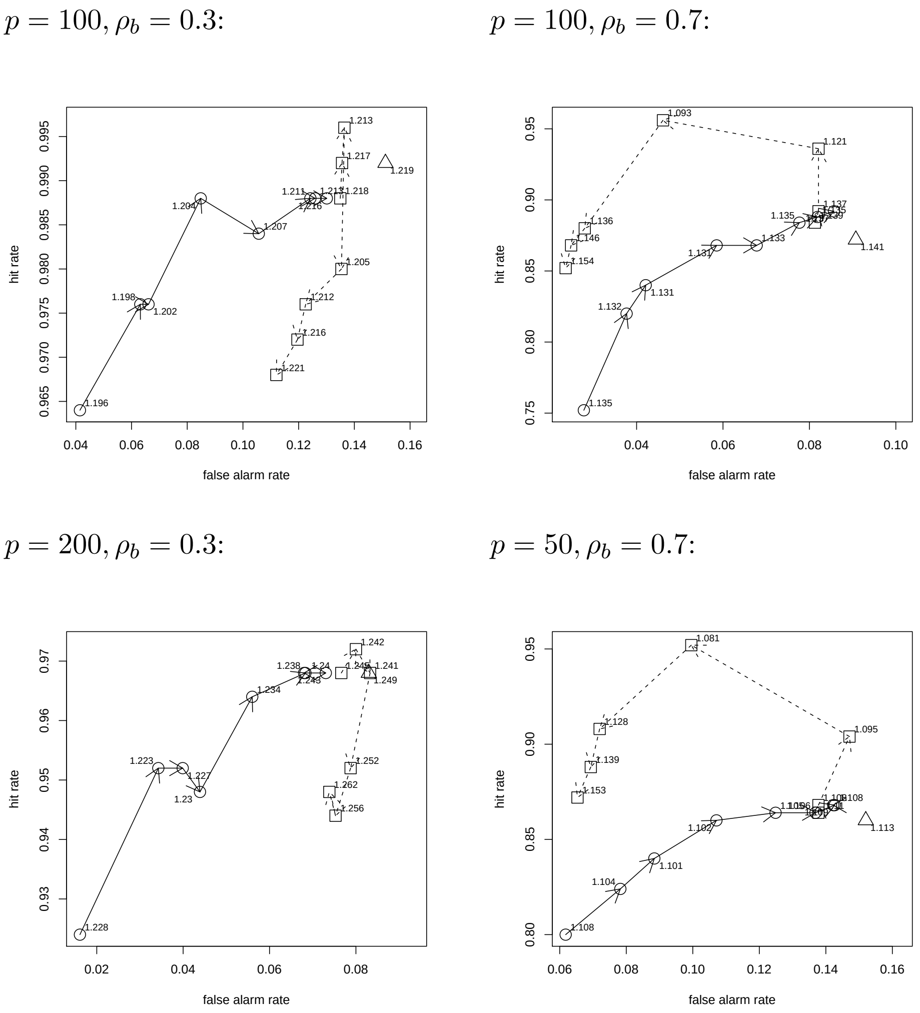
Figure 5 shows the hit rates and false alarm rates for componentwise boosted ridge regression (circles), elastic net (squares) and the Lasso (triangle) for the data underlying Figure 4. While the Lasso has only one parameter which is selected for optimal prediction performance, the componentwise approach and elastic net have two parameters, a penalty parameter and the number of steps. For evaluation of prediction performance we used a fixed penalty for the componentwise approach and the optimal penalty (with respect to prediction performance) for the elastic net procedure. For the investigation of their properties with respect to identification of influential variables we vary the penalty parameters (the number of steps still being chosen for optimal performance), thus resulting in a sequence of fits. Plotting the hit rates and false alarm rates of these fits leads to the pseudo-ROC-curves shown in Figure 5. We call them "pseudo" curves since they are not necessarily monotone!

It is seen that for a large number of covariates and a medium level of correlation (bottom left panel) the componentwise approach comes close to dominating the other procedures. While for a smaller number of variables and medium level of correlation (top left panel) higher hit rates can be achieved by using the elastic net procedure the componentwise approach still is the only procedure which allows for a trade-off of hit rate and false alarm rate (and therefore selection of small false alarm rates) by variation of the penalty parameter. In this case for the elastic net the false alarm rate hardly changes.

For the examples with high correlations between covariates (right panels) there is a clear advantage for the elastic net procedure. With penalty parameters going toward zero elastic net comes close to a Lasso solution (as it should, based on its theory). This is also where the elastic net pseudo-ROC curve (for small penalties) and the curve of the componentwise approach (for large penalties) coincide. The differences seen between elastic net/componentwise solutions with small/large penalties and the Lasso solution might result from the different type of tuning parameter (number of steps vs. constraint on the parameter vector).

### 5.2 Binary response

In order to evaluate generalized boosted ridge regression and generalized componentwise boosted ridge regression for the non-metric response case we compare them to a generalized variant of the Lasso, which is obtained by using iteratively re-weighted least



$p = 100, \rho_b = 0.3$:

$p = 100, \rho_b = 0.7$:

$p = 200, \rho_b = 0.3$:

$p = 50, \rho_b = 0.7$:

Figure 5: Pseudo-ROC curves for the identification of influential and non-influential covariates with metric response data for PartBoostR (circles) and elastic net (squares) with varying penalty parameters (and the optimal number of steps) and for the Lasso (triangle). Arrows indicate increasing penalty parameters. The numbers give the mean squared error of prediction for the respective estimates.

squares with a pseudo-response where weighted least-squares estimation is replaced by a weighted Lasso estimate (see documentation of the "lasso2" package Lokhorst, 1999). The Lasso constraint parameter and the number of steps for (componentwise) boosted ridge regression are determined by cross-validation and we compare the resulting per-

Table 1: Mean deviance (of prediction) for binary response data with varying number of predictors $p$ and correlation $\rho_b$ for a generalized linear model including only an intercept term (base), generalized Lasso (with optimal constraint and constraint selected by cross-validation), generalized boosted ridge regression (GenBoostR) and generalized componentwise boosted ridge regression (GenPartBoostR) (with optimal number of steps and number of steps selected by AIC, BIC or cross-validation (cv)). The best result obtainable by data selected tuning parameters is marked in boldface for each example.

| $p$ | $\rho_b$ | base | Lasso | | GenBoostR | | | GenPartBoostR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | opt | cv | opt | AIC | cv | opt | AIC | BIC | cv |
| 10 | 0 | 1.397 | 0.875 | 0.896 | 0.885 | 0.900 | 0.896 | 0.873 | 0.904 (4) | 0.899 | **0.893** |
| | 0.3 | 1.400 | 0.873 | 0.887 | 0.874 | 0.897 | 0.886 | 0.871 | 0.894 (2) | 0.900 | **0.886** |
| | 0.7 | 1.390 | 0.845 | 0.856 | 0.834 | 0.849 | **0.844** | 0.844 | 0.858 (0) | 0.858 | 0.854 |
| 50 | 0 | 1.398 | 0.935 | 0.963 | 1.098 | 1.371 | 1.143 | 0.931 | 1.058 (7) | **0.953** | 0.969 |
| | 0.3 | 1.391 | 0.931 | 0.998 | 1.035 | 1.318 | 1.068 | 0.921 | 1.093 (6) | **0.936** | 0.985 |
| | 0.7 | 1.395 | 0.888 | 0.910 | 0.924 | 1.034 | 0.937 | 0.883 | 1.051 (11) | **0.891** | 0.905 |
| 100 | 0 | 1.395 | 1.019 | 1.042 | 1.241 | 1.281 | 1.265 | 1.014 | 1.296 (11) | **1.029** | 1.064 |
| | 0.3 | 1.394 | 0.991 | 1.011 | 1.159 | 1.234 | 1.187 | 0.981 | 1.256 (14) | **0.991** | 1.014 |
| | 0.7 | 1.395 | 0.925 | 0.962 | 0.996 | 1.052 | 1.057 | 0.918 | 1.209 (16) | **0.936** | 0.986 |
| 200 | 0 | 1.395 | 1.075 | 1.110 | 1.307 | 1.350 | 1.330 | 1.071 | 1.332 (1) | **1.095** | 1.119 |
| | 0.3 | 1.396 | 1.010 | 1.049 | 1.261 | 1.327 | 1.283 | 1.002 | 1.232 (4) | **1.013** | 1.051 |
| | 0.7 | 1.402 | 0.950 | 0.972 | 1.113 | 1.164 | 1.141 | 0.944 | 1.278 (12) | **0.963** | 0.976 |

formance to that obtained with optimal parameter values. For the binary response examples given in the following the AIC and the BIC (see Section 4) is available as an additional criterion for (componentwise) boosted ridge regression.

Table 1 gives the mean deviance for binary response data generated from (3) with $n = 100$ and a varying number of variables and correlation for 20 repetitions per example. When using the AIC as a stopping criterion in several instances (number indicated in parentheses) no minimum within the range of 500 boosting steps could be found and so effectively the maximum number is used. It is seen that for a small number of variables and high correlation between covariates ($p = 10, \rho_b = 0.7$)(generalized) ridge regression and componentwise boosted ridge regression perform similar, for all other examples the componentwise approach is ahead. This parallels the findings from the continuous response examples. While it seems that there is a slight advantage of the componentwise approach over the Lasso, their performance (with optimal parameters as well as with parameters selected by cross-validation) is very similar. Selection of the number of boosting steps by AIC seems to work well only for a small number of variables, as can be seen e.g. when comparing to the cross-validation results for the componentwise approach. For a larger number of covariates the use of AIC for the selection of the number of boosting steps seems to be less efficient. In contrast BIC seems to perform quite well for a moderate to large number of covariates while being suboptimal for a smaller number of predictors.

Table 2 shows the hit rates/false alarm rates obtained from the the componentwise approach and from the Lasso when using optimal parameters (number of boosting steps for GenPartBoostR and the constraint on the parameter vector for the Lasso) as well as for parameters chosen by cross-validation and BIC (for the componentwise approach). It is seen that with optimal parameters application of the componentwise approach can result in smaller false alarm rates compared to the Lasso while maintaining a competitive hit rate. This is similar to the continuous response examples and therefore the trade-off between hit rate and false alarm by selection of the penalty parameter seems to be feasible. Using BIC as a criterion for the selection of the number of boosting steps the combination of hit rate and false alarm rate obtained with componentwise boosted ridge regression even dominates the Lasso results in several instances (printed in boldface).

# 6 Application

We illustrate the application of (componentwise) boosted ridge regression with real data. The data are from 344 admissions at a psychiatric hospital with a specific diagnosis

Table 2: Hit rates/false alarm rates for identification of influential covariates with binary response data. Combinations of hit rate and false alarm rate obtained by cross-validation that dominate all other procedures for an example are printed in boldface.

| $p$ | $\rho_b$ | Lasso | | GenPartBoostR | | |
|---|---|---|---|---|---|---|
| | | opt | cv | opt | BIC | cv |
| 10 | 0 | 1/0.650 | 1/0.660 | 1/0.620 | 0.99/0.350 | 1/0.640 |
| | 0.3 | 0.98/0.560 | 0.98/0.530 | 0.98/0.520 | 0.96/0.310 | **0.98/0.510** |
| | 0.7 | 0.91/0.540 | 0.89/0.400 | 0.92/0.530 | 0.87/0.350 | 0.88/0.410 |
| 50 | 0 | 0.99/0.287 | 1/0.290 | 0.99/0.250 | 0.99/0.156 | 1/0.260 |
| | 0.3 | 0.97/0.226 | 0.98/0.223 | 0.97/0.191 | **0.98/0.144** | 0.98/0.203 |
| | 0.7 | 0.87/0.158 | 0.85/0.131 | 0.87/0.133 | **0.86/0.108** | 0.86/0.124 |
| 100 | 0 | 0.94/0.161 | 0.93/0.174 | 0.94/0.136 | **0.93/0.125** | 0.92/0.175 |
| | 0.3 | 0.93/0.153 | 0.95/0.131 | 0.96/0.121 | 0.94/0.109 | 0.95/0.134 |
| | 0.7 | 0.79/0.109 | 0.77/0.098 | 0.81/0.089 | **0.80/0.090** | 0.76/0.102 |
| 200 | 0 | 0.89/0.088 | 0.90/0.097 | 0.89/0.067 | 0.88/0.085 | 0.89/0.088 |
| | 0.3 | 0.95/0.096 | 0.94/0.096 | 0.95/0.073 | **0.95/0.079** | 0.93/0.089 |
| | 0.7 | 0.85/0.062 | 0.84/0.071 | 0.83/0.053 | 0.86/0.070 | 0.85/0.068 |

within a timeframe of eight years. The (binary) response variable to be investigated is whether treatment is aborted by the patient against physicians' advice (about 55% for this diagnosis). There are five metric and 16 categorical variables from routine documentation at admission available for prediction. This encompasses socio-demographic variables (age, sex, education, employment, etc.) as well as clinical information (level of functioning, suicidal behavior, medical history, etc.). After re-coding each categorical variable into several 0/1-variables for easier inclusion into a linear model there is a total of 101 predictors. The clinic's interest is not primarily exact prediction but identification of informative variables that allow for an early intervention to prevent patients with high risk from aborting treatment. Nevertheless we divide the data into 270 admissions from the first six year interval for model building and the 74 admissions of the last two years for model validation. The baseline prediction error on the latter data using an intercept-only model is 0.446.

In a first step we apply componentwise boosted ridge regression and the Lasso with the full set of 101 predictors. From the physicians' view one variable is of special interest and should be incorporated in the analysis. This is the secondary diagnosis assigned to the patient (taking either the value "none" or one of six diagnosis groups). For the Lasso we can either include the predictors coding for the secondary diagnosis without restriction or impose the same bound used for the other variables. We use the latter option, although this leads to estimates for only two of the secondary diagnosis groups, because the other predictors are excluded. With componentwise boosted ridge regression there is the option of making the secondary diagnosis predictors mandatory members of the candidate sets $V_m^{(j)}$, i.e. there will be estimates for them in any case. For a further reduction of the complexity we use a special structure for the penalty matrix $P$ with larger penalty for the mandatory elements of the candidate sets. This option of using a customized penalization structure is another distinct feature of the componentwise approach in comparison to the Lasso.

Table 3 shows the number of variables with non-zero parameter estimates, the approximate degrees of freedom and the prediction error on the test data for the models resulting from componentwise boosted ridge regression with varying penalty parameter (with BIC) and the Lasso (with cross-validation). Note that the reason why the Lasso uses less variables compared to the componentwise approach is the use of the mandatory parameters with the latter. It is seen that the performance of the componentwise approach is slightly superior to the Lasso. What is more interesting is that with varying penalty parameter the number of variables used for optimal performance also varies. This feature of the componentwise approach — using the penalty parameter to control the hit rate/false alarm rate (and thereby the number of variables in the model) while maintaining performance — has already been illustrated in Section 5, but it is assuring that it can also be found with real data. It should be noted that the number of vari-

Table 3: Number of variables with non-zero estimates, approximate degrees of freedom and prediction error on the test set for the binary response example with 101 predictors using componentwise boosted ridge regression with various penalty parameters penalty and the Lasso. The number of boosting steps is selected by BIC and the Lasso bound is selected by cross-validation.

| | penalty | number of variables | degrees of freedom | prediction error |
|---|---|---|---|---|
| the Lasso | - | 13 | - | 0.392 |
| PartBoostR | 500 | 15 | 3.545 | 0.378 |
| | 1000 | 15 | 3.124 | 0.392 |
| | 2000 | 18 | 3.363 | 0.392 |
| | 5000 | 19 | 3.292 | 0.378 |
| | 10000 | 19 | 3.292 | 0.378 |

ables is not proportional to model complexity (represented by the approximate degrees of freedom). It seems that while the degrees of freedom approximately stay the same they are allocated to the covariates in a different way.

We already showed the coefficient build-up for a metric response example in Section 3.1. To obtain illustrative data for the binary response example we examine the predictors selected by the componentwise approach in the analysis above, combine several of those into new variables and add additional variables based on subject matter considerations. The variables used in the following are age, number of previous admissions ("noupto"), cumulative length of stay ("losupto") and the 0/1-variables indicating a previous ambulatory treatment ("prevamb"), no second diagnosis ("nosec"), second diagnosis "personality disorder" ("persdis"), somatic problems ("somprobl"), homelessness ("homeless") and joblessness ("jobless"). In contrast to the previous example with respect to the second diagnosis only the two most important group variables are used to keep the coefficient build-up graphs simple. Those two variables ("nosec" and "persdis") again are mandatory members of the response set, but this time no penalty is applied to

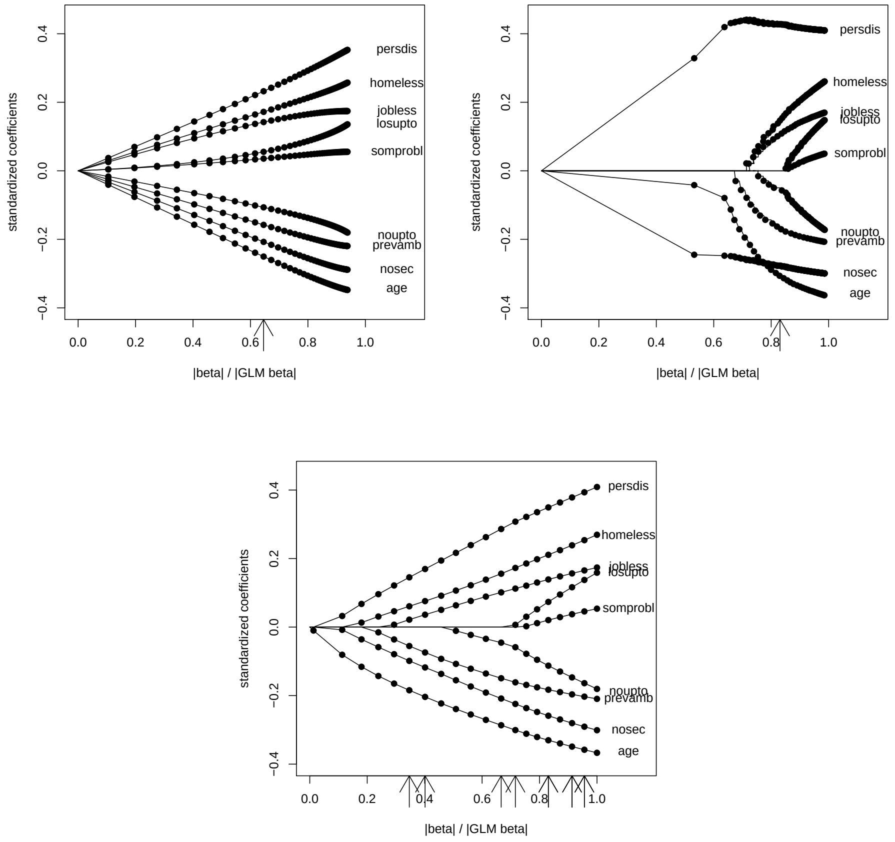


Figure 6: Coefficient build-up for example data (boosted ridge regression: upper left; componentwise boosted ridge regression: upper right; the Lasso: bottom panel).

their estimates. This illustrates the effect of augmenting an unpenalized model with few mandatory variables with optional predictors. The top right panel of Figure 6 shows the coefficient build-up in the course of the boosting steps for componentwise boosted ridge regression contrasted with boosted ridge regression (top left) and the Lasso (bottom panel). The arrows indicate the number of steps chosen by AIC (for (componentwise) boosted ridge regression) and 10-fold cross-validation (for the Lasso; repeated for 10 times). It can be seen that while for the optional predictors the componentwise approach results in a build-up scheme similar to the Lasso, the mandatory components introduce a very different structure. One interesting feature is the slow decrease of the estimate for "persdis" beginning with boosting step 8. This indicates some overshooting of the initial estimate that is corrected when additional predictors are included.

## 7 Concluding remarks

We investigated the use of boosting for the fitting of linear models. The first estimate we considered was the boosted ridge estimator. Although the boosted ridge estimator yields shrinkage that differs from the simple ridge estimator in practice the performance does not differ strongly. Strongly improved estimates are obtained for partial boosting, in particular for small correlation between covariates. The performance of BoostPartR turns out to be similar to that of stagewise regression procedures which may be seen as competitors, in particular for large numbers of covariates. In empirical comparisons it is seen that the componentwise approach is competitive to the Lasso and the elastic net procedures in terms of prediction performance as well as with respect to the identification of influential covariates. The elastic net procedure (not surprisingly) performs best in situations with highly correlated predictors. For examples with less correlation componentwise boosted ridge regression is competitive not only in terms of prediction performance but also with respect to covariate identification. This seems to hold not only for the continuous response case but also for binary responses. An additional feature of the componentwise approach is that the trade-off between hit rate/false alarm rate can be controlled over a broad range by selecting the penalty parameter. Especially in situations where a small false alarm rate is wanted (e.g. for obtaining parsimonious models) this is a distinct advantage over the other procedures.

## Appendix

*Proof (Proposition 1).* One obtains

$$cov(\hat{\mu}_m) = \sigma^2 U(I_n - (I_n - \widetilde{D})^{m+1})^2 U^{\mathrm{T}}$$

immediately from $\hat{\mu}_{(m)} = H_m y$.

The bias is obtained from

$$\begin{aligned} b &= E(\mu - \mu_m) = (I_n - H_m)\mu \\ &= (I_n - (I_n - (I_n - U\widetilde{D}U^{\mathrm{T}})^{m+1}))\mu \\ &= (I_n - U\widetilde{D}U^{\mathrm{T}})^{m+1}\mu = U(I_n - \widetilde{D})^{m+1}U^{\mathrm{T}}\mu \end{aligned}$$

Therefore one obtains the squared bias

$$\begin{aligned} b^{\mathrm{T}}b &= \mu^{\mathrm{T}}\, U(I_n - \widetilde{D})^{m+1}U^{\mathrm{T}}\, U(I_n - \widetilde{D})^{m+1}U^{\mathrm{T}}\, \mu \\ &= \mu^{\mathrm{T}}\, U\, (I_n - \widetilde{D})^{2(m+1)}\, U^{\mathrm{T}}\, \mu \\ &= \mu^{\mathrm{T}}\, U\; Diag((1 - \tilde{d}_1^2)^{2m+2}, \ldots, (1 - \tilde{d}_p^2)^{2m+2})\; U^{\mathrm{T}}\, \mu \end{aligned}$$

and therefore

$$\begin{aligned} trace\; cov(H_m y) &= trace(\sigma^2\; U(I_n - (I_n - \widetilde{D})^{m+1})^2\; U^{\mathrm{T}}) \\ &= trace(\sigma^2 \sum_{j=1}^{p} u_j\; (1 - (1 - \tilde{d}_j^2)^{m+1})^2\; u_j^T) \\ &= \sum_{j=1}^{r} \sigma^2 (1 - (1 - \frac{d_j^2}{d_j^2 + \lambda})^{m+1})^2 \end{aligned}$$

yielding

$$\begin{aligned} MSE(BoostR(m)) &= \frac{1}{n}\; (trace\; cov(H_m y) + b^{\mathrm{T}}b) \\ &= \frac{1}{n}\; (\sum_{j=1}^{r} \sigma^2 (1 - (1 - \tilde{d}_j^2)^{m+1})^2 \\ &+ \;\mu^{\mathrm{T}}\, U\; Diag((1 - \tilde{d}_1^2)^{2m+2}, \ldots, (1 - \tilde{d}_p^2)^{2m+2})\; U^{\mathrm{T}}\, \mu) \quad \square \end{aligned}$$

*Proof (Proposition 2).* The MSE of boosted ridge ($\lambda > 0$) is given by

$$MSE(BoostR_\lambda(m)) = \frac{1}{n} \sum_{j=1}^{p} \{\sigma^2 (1 - (1 - \tilde{d}_j^2)^{m+1})^2 + c_j (1 - \tilde{d}_j^2)^{2m+2})\}$$

where $\tilde{d}_j^2 = d_j^2/(d_j^2 + \lambda)$, and $c_j = \mu^T u_j u_j^T \mu = \|\mu^T u_j\|$ depends only on the underlying model.

The MSE of the least squares estimate is given by $MSE(ML) = \frac{r}{n}\sigma^2$.

Since $\tilde{d}_j^2 = c_j = 0$ for $j > r$ and $0 < \tilde{d}_j^2 \leq 1$ for $j \leq r$ (a) follows immediately. In addition one obtains $MSE(BoostR_\lambda(m)) \leq MSE(ML)$ if

$$\sum_{j=1}^{r} (1-(1-\tilde{d}_j^2)^{m+1})^2 + \frac{c_j}{\sigma^2}(1-\tilde{d}_j^2)^{2m+2} \leq r,$$

which is equivalent to

$$\sum_{j=1}^{r} \left\{(1-\tilde{d}_j^2)^{2m+2}(\frac{c_j}{\sigma^2}+1) + 1 - 2(1-\tilde{d}_j^2)^{m+1}\right\} \leq r,$$

and

$$\sum_{j=1}^{r} 2(1-\tilde{d}_j^2)^{m+1} - (1-\tilde{d}_j^2)^{2(m+1)}(\frac{c_j}{\sigma^2}+1) \geq 0.$$

It is enough to find $\lambda$, $m$ such that

$$\sum_{j=1}^{r} 2(1-\tilde{d}_j^2)^{m+1}\left(2-(1-\tilde{d}_j^2)^{m+1}(\frac{c_j}{\sigma^2}+1)\right) \geq 0$$

Since $\lambda > 0$ one can choose $m$ so large that $(1-\tilde{d}_j^2)^{m+1}$ becomes arbitrarily small. Therefore (b) holds. □

*Proof (Proposition 3).* In the $m$th iteration (after $V_m$ has been selected) the update is given by

$$\begin{aligned}\hat{\beta}_{V_m}^R &= F_{p,V_m}^{-1} s_{p,V_m} \\ &= (X_{V_m}^T W_m X_{V_m} + \lambda I_p)^{-1} X_{V_m} W_m D_m^{-1}(y - \hat{\mu}_{(m-1)})\end{aligned}$$

where $W_m = W(\hat{\eta}_{(m-1)})$, $D_m = D(\hat{\eta}_{(m-1)})$ are evaluated at $\hat{\eta}_{(m-1)}$.

One has

$$\begin{aligned}\hat{\eta}_{(m)} - \hat{\eta}_{(m-1)} &= X\hat{\beta}_{(m-1)} + X_{V_m}\hat{\beta}_{V_m} - X\hat{\beta}_{(m-1)} = X_{V_m}\hat{\beta}_{V_m} \\ &= X_{V_m}(X_{V_m}^T W_m X_{V_m} + \lambda I_p)^{-1} X_{V_m} W_m D_m^{-1}(y - \hat{\mu}_{(m-1)}).\end{aligned}$$

By using Taylor approximation of first order $h(\hat{\eta}) \approx h(\eta) + (\partial h(\eta)/\partial \eta^T)(\hat{\eta} - \eta)$ one obtains

$$\begin{aligned}\hat{\mu}_{(m)} &\approx \hat{\mu}_{(m-1)} + D_m(\hat{\eta}_{(m)} - \hat{\eta}_{(m-1)}) = \hat{\mu}_{(m-1)} + D_m X_{V_m}\hat{\beta}_{V_m}^R \\ &= \hat{\mu}_{(m-1)} + D_m X_{V_m}(X_{V_m}^T W_m X_{V_m} + \lambda I_p)^{-1} X_{V_m}^T W_m D_m^{-1}(y - \hat{\mu}_{(m-1)}).\end{aligned}$$

and therefore

$$D_m^{-1}(\hat{\mu}_{(m)} - \hat{\mu}_{(m-1)}) \approx X_{V_m}(X_{V_m}^T W_m X_{V_m} + \lambda I_p)^{-1} X_{V_m} W_m D_m^{-1}(y - \hat{\mu}_{(m-1)}).$$

Multiplication with $W_m^{1/2}$ and using $W_m^{1/2} D_m^{-1} = \Sigma_m^{-1/2} = diag(\sigma_{(m-1),1}, \dots, \sigma_{(m-1),n})$ yields

$$\Sigma_m^{-1/2}(\hat{\mu}_{(m)} - \hat{\mu}_{(m-1)}) \approx \tilde{H}_m \Sigma_m^{-1/2}(y - \hat{\mu}_{(m-2)}).$$

where $\tilde{H}_m = W^{1/2} X_{V_m}(X_{V_m}^T W_m X_{V_m} + \lambda I_p)^{-1} X_{V_m} W^{1/2}$. Defining $M_m = \Sigma_m^{1/2} \tilde{H}_m \Sigma_m^{-1/2}$ yields the approximation

$$\begin{aligned}
\hat{\mu}_{(m)} &\approx \hat{\mu}_{(m-1)} + M_m(y - \hat{\mu}_{(m-1)}) \\
&= \hat{\mu}_{(m-1)} + M_m(y - \hat{\mu}_{(m-2)}) - (\hat{\mu}_{(m-1)} - \hat{\mu}_{(m-2)}) \\
&\approx \hat{\mu}_{(m-1)} + M_m((y - \hat{\mu}_{(m-2)}) - M_{m-1}(y - \hat{\mu}_{(m-2)})) \\
&= \hat{\mu}_{(m-1)} + M_m(I_n - M_{m-1})(y - \hat{\mu}_{(m-2)}).
\end{aligned}$$

With starting value $\hat{\mu}_{(0)} = M_0 y$ one obtains

$$\hat{\mu}_{(1)} \approx \hat{\mu}_{(0)} + M_1(y - \hat{\mu}_{(0)}) = M_0 y + M_1(I_n - M_0)y$$

and more general

$$\hat{\mu}_{(m)} \approx H_m y$$

where

$$H_m = \sum_{j=0}^{m} M_j \prod_{i=0}^{j-1}(I_n - M_0). \qquad \square$$

**Acknowledgements**

We gratefully acknowledge support from Deutsche Forschungsgemeinschaft (Project C4, SFB 386 Statistical Analysis of Discrete Structures).

# References

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11:1493–1517.

Bühlmann, P. (2005). Boosting for high-dimensional data. *The Annals of Statistics*, (to appear).

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

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

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

Friedman, J. H., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Jounral of Computational and Graphical Statistics*, 7(3):397–416.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

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

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Le Cessie, S. and van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

Lokhorst, J. (1999). The lasso and generalised linear models. Honors Project.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5:197–227.

Seber, G. A. F. (1977). *Linear Regression Analysis*. Wiley, New York.

Stamney, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, ii: Radical prostatectomy treated patients. *Journal of Urology*, 16:1076–1083.

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

Tutz, G. and Binder, H. (2004). Generalized additive modelling with implicit variable selection by likelihood based boosting. Discussion Paper 401, SFB 386, Ludwig Maximilians University Munich.

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