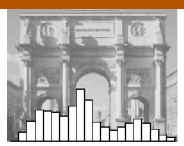
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK



Matthias Schmid, Sergej Potapov, Annette Pfahlberg & Torsten Hothorn

Estimation and Regularization Techniques for Regression Models with Multidimensional Prediction Functions

Technical Report Number 042, 2008
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Estimation and Regularization Techniques for Regression Models with Multidimensional Prediction Functions

Matthias Schmid[1], Sergej Potapov[1],
Annette Pfahlberg[1], Torsten Hothorn[2]

Abstract

Boosting is one of the most important methods for fitting regression models and building prediction rules from high-dimensional data. A notable feature of boosting is that the technique has a built-in mechanism for shrinking coefficient estimates and variable selection. This regularization mechanism makes boosting a suitable method for analyzing data characterized by small sample sizes and large numbers of predictors. We extend the existing methodology by developing a boosting method for prediction functions with multiple components. Such multidimensional functions occur in many types of statistical models, for example in count data models and in models involving outcome variables with a mixture distribution. As will be demonstrated, the new algorithm is suitable for both the estimation of the prediction function and regularization of the estimates. In addition, nuisance parameters can be estimated simultaneously with the prediction function.

Keywords: Gradient boosting, multidimensional prediction function, scale parameter estimation, variable selection, count data model, clinical predictors.

1 Introduction

A common problem in statistical research is the development of model fitting and prediction techniques for the analysis of high-dimensional data. High-dimensional data sets, which are characterized by relatively small sample sizes and large numbers of variables, arise in many fields of modern research. Most notably, advances in genomic research have led to large sets of gene expression data where sample sizes are considerably smaller than the number of gene expression measurements (Golub et al. 1999, Dudoit et al. 2002). A consequence of this "$p > n$" situation is that standard techniques for prediction and model fitting (such as maximum likelihood estimation) become infeasible. Moreover, high-dimensional data sets usually involve the problem of separating noise from information, i.e., of selecting a small number of relevant predictors from the full set of variables.

[1]Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Waldstraße 6, 91054 Erlangen, Germany

[2]Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstraße 33, 80539 München, Germany

In a regression framework, the problem of analyzing high-dimensional data can be formulated as follows: Consider a data set containing the values of an outcome variable $\boldsymbol{Y}$ and predictor variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$. Although $\boldsymbol{Y}$ will be one-dimensional in most applications, we explicitly allow for multidimensional outcome variables. The objective is to model the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)^\top$, and to obtain an "optimal" prediction of $\boldsymbol{Y}$ given $\boldsymbol{X}$. Usually, this is accomplished by optimizing an objective function $\rho(\boldsymbol{Y}, f, \sigma) \in \mathbb{R}$ over a prediction function f (depending on $\boldsymbol{X}$) and a set of scale parameters (denoted by σ). Linear regression with a continuous outcome variable $\boldsymbol{Y} \in \mathbb{R}$ is a well-known example of this approach: Here, ρ corresponds to the least squares objective function while f is a parametric (linear) function of $\boldsymbol{X}$, and $\sigma \in \mathbb{R}^+$ is the residual variance.

In order to address the issue of analyzing high-dimensional data sets, a variety of regression techniques has been developed over the past years (see, e.g., Hastie et al. 2003). Many of these techniques are characterized by a built-in mechanism for "regularization", which means that shrinkage of coefficient estimates or selection of relevant predictors is carried out simultaneously with the estimation of the model parameters. Both shrinkage and variable selection will typically improve prediction accuracy: In case of shrinkage, coefficient estimates tend to have a slightly increased bias but a decreased variance, while in case of variable selection, overfitting the data is avoided by selecting the most "informative" predictors only.

Important examples of recently developed regularization techniques are boosting (which will be considered in this paper) and L_1 penalized estimation. *Boosting* (Breiman 1998, 1999, Friedman et al. 2000, Friedman 2001) is an iterative method for obtaining statistical model estimates via gradient descent techniques. A key feature of boosting is that the procedure can be modified such that variable selection is carried out in each iteration (Bühlmann and Yu 2003, Bühlmann 2006). As a result, the final boosting fit typically depends on only a small subset of predictor variables but can still be interpreted as the fit of a regression model. L_1 *penalized estimation* techniques have been developed for regression models with a linear prediction function. Due to the structure of the L_1 penalty, a number of coefficient estimates will typically become zero, so that the procedure implicitly results in a selection of the most informative predictor variables. The most important examples of L_1 penalized techniques are the Lasso and its extensions (Tibshirani 1996, Tibshirani et al. 2005, Zou 2006, Yuan and Lin 2006), SCAD procedures (Fan and Li 2001) and the Elastic Net methodology (being a combination of L_1 and L_2 penalized regression, see Zou and Hastie 2005). By introducing the LARS algorithm for linear prediction functions, Efron et al. (2004) have embedded boosting and L_1 penalized techniques into a more general framework (LARS will, however, not be considered in this paper). Both boosting and L_1 penalized estimation techniques can be applied to a large variety of statistical problems, such as regression, classification and time-to-event analysis (Bühlmann and Hothorn 2007, Park and Hastie 2007). Besides being computationally efficient, the techniques are competitive with methods based on a separation of variable selection and the model fitting process (see, e.g., Segal 2006).

A limitation of classical boosting and L_1 penalized estimation approaches is that the techniques are designed for statistical problems involving a *one-dimensional* prediction function only. In

fact, boosting and L_1 penalized estimation are suitable for fitting many common statistical models, such as linear or logistic regression. However, there is a variety of important statistical problems that cannot be reduced to estimating a one-dimensional prediction function only. This is particularly true when scale parameters or nuisance parameters have to be estimated simultaneously with the prediction function, or when the prediction function itself depends on multiple components. Typical examples of such multidimensional estimation problems are:

(a) *Classification with multiple outcome categories.* Regressing outcome variables with a multinomial distribution on a set of predictor variables is a natural extension of the binary classification problem. In the setting of a multinomial logit model, each of the outcome categories is associated with a separate component of the prediction function. Thus, if there is a total number of K possible outcome categories, a K-dimensional prediction function has to be estimated. Friedman et al. (2000) have addressed this problem by constructing a boosting algorithm for multiclass prediction (see also Hastie et al. 2003, Sections 10.10.2 and 10.10.3).

(b) *Regression models for count data.* Apart from the classical Poisson model, count data models are typically used for addressing problems such as overdispersion or excessive amounts of zero counts. A typical example in this context is negative binomial regression, where the prediction function has to be estimated simultaneously with a scale parameter used to model overdispersion. If excessive amounts of zero counts have to be taken into account, it is common to use zero-inflated Poisson or negative binomial models (see Hilbe 2007). With models of this type, the outcome variable is assumed to be a mixture of a zero-generating (Bernoulli) process and a counting process. As a consequence, using zero-inflated Poisson or negative binomial models involves the estimation of a two-dimensional prediction function (where the first component of the prediction function is used to model the zero-generating process and the second component is used to model the count data process). It is important to note that each of the two components may depend on different sets of predictor variables. Furthermore, fitting zero-inflated negative binomial models involves the estimation of an additional scale parameter, where both the two-dimensional prediction function and the scale parameter have to be estimated simultaneously.

(c) *Clinical predictors in cancer research.* Predicting outcomes such as "time to disease" or "future disease status" is a common problem in cancer research. In many situations, there are well-established predictors for the outcome variable(s) under consideration (such as the International Prognostic Index, see International Non-Hodgkin's Lymphoma Prognostic Factors Project 1993). At the same time, advances in genomic research have lead to new predictors based on gene expression measurements (see, e.g., van't Veer et al. 2002, Bullinger et al. 2004). It is an obvious question whether the performance of cancer prediction can be improved by combining gene-based and traditional clinical predictors. Clearly, a combined model involves prediction functions depending on two components, where one component depends on a set of clinical predictor variables and the other component depends on gene expression measurements. Since inclusion of the clinical component is often mandatory (for both practical and conceptual reasons), only the gene-based component of the prediction function requires variable selection (i.e., the selection of a small number of relevant genes).

Obviously, all examples described above are special cases of a more general estimation problem involving predictors with multiple components. We will address this problem by developing a boosting algorithm for multidimensional prediction functions. The proposed algorithm is based on the classical gradient boosting method introduced by Friedman (2001) but is modified such that both parameter estimation and variable selection can be carried out in each component of the multidimensional prediction function. Instead of "descending" the gradient only in one direction, the algorithm computes partial derivatives of the objective function (with respect to the various components of the prediction function). In a next step, the algorithm cycles through the partial derivatives, where each component of the prediction function is successively updated in the course of the cycle. Similar to the original boosting idea, this procedure can be modified such that variable selection is carried out in each step of the cycle. If necessary, updates of scale parameters can be obtained at the end of the cycle. This is accomplished by using the current value of the prediction function as an offset value.

As we will demonstrate, the new algorithm constitutes a flexible approach to model fitting and prediction in multidimensional settings. Moreover, the algorithm shares the favorable properties of the classical boosting approach when it comes to efficiency and prediction accuracy. In the special case of a one-dimensional prediction function, the new approach coincides with the original boosting algorithm proposed by Friedman (2001). In addition, it generalizes the work of Schmid and Hothorn (2008b) who developed a boosting algorithm for parametric survival models with a scale parameter. In case of a multinomial logit model, there is a direct correspondence between the new algorithm and the multiclass procedure suggested by Friedman et al. (2000).

The rest of the paper is organized as follows: In Section 2, the new algorithm is presented in detail, along with a number of technical details involved in choosing appropriate tuning parameters. The characteristics of the algorithm are demonstrated in Section 3, where two examples from epidemiological and clinical research are discussed. A summary of the paper is given in Section 4.

2 Boosting with multidimensional prediction functions

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a set of independent realizations of the random variable $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{X}$ is a p-dimensional vector of predictor variables and $\boldsymbol{Y}$ is a (possibly multidimensional) outcome variable. Define $X := (X_1, \ldots, X_n)$ and $Y := (Y_1, \ldots, Y_n)$. The objective is to estimate the K-dimensional prediction function $f^* \in \mathbb{R}^K$ and the L-dimensional set of scale parameters $\sigma^* \in \mathbb{R}^L$, which are defined by

$$(f^*, \sigma^*) = (f_1^*, \ldots, f_K^*, \sigma_1^*, \ldots, \sigma_L^*) := \underset{f,\sigma}{\operatorname{argmin}} \, \mathbb{E}_{\boldsymbol{Y},\boldsymbol{X}}\big[\rho(\boldsymbol{Y}, f(\boldsymbol{X}), \sigma)\big] \; . \tag{1}$$

The objective function (or "loss function") ρ is assumed to be differentiable with respect to each of the components of $f = (f_1, \ldots, f_K)$.

Usually, in a boosting framework, f^* and σ^* are estimated by minimizing the empirical risk $\sum_{i=1}^{n} \rho(Y_i, f(X_i), \sigma)$ over f and $\sigma = (\sigma_1, \ldots, \sigma_L)$. We introduce the following multidimensional extension of the gradient boosting approach developed by Friedman (2001):

1. Initialize the n-dimensional vectors $\hat{f}_1^{[0]}, \ldots, \hat{f}_K^{[0]}$ with offset values, e.g., $\hat{f}_1^{[0]} = \mathbf{0}, \ldots, \hat{f}_K^{[0]} = \mathbf{0}$. Further initialize the one-dimensional scale parameter estimates $\hat{\sigma}_1^{[0]}, \ldots, \hat{\sigma}_L^{[0]}$ with offset values, e.g., $\hat{\sigma}_1^{[0]} = 1, \ldots, \hat{\sigma}_K^{[0]} = 1$. (Alternatively, the maximum likelihood estimates corresponding to the unconditional distribution of $\boldsymbol{Y}$ could be used as offset values.)

2. For each of the K components of f specify a *base-learner*, i.e., a regression estimator with one input variable and one output variable. Set $m = 0$.

3. Increase m by 1.

4. (a) Set $k = 0$.

   (b) Increase k by 1. Compute the negative partial derivative $-\frac{\partial \rho}{\partial f_k}$ and evaluate at

$$\hat{f}^{[m-1]}(X_i) = \left(\hat{f}_1^{[m-1]}(X_i), \ldots, \hat{f}_K^{[m-1]}(X_i)\right) , \tag{2}$$

$$\hat{\sigma}^{[m-1]} = \left(\hat{\sigma}_1^{[m-1]}, \ldots, \hat{\sigma}_L^{[m-1]}\right) , \; i = 1, \ldots, n. \tag{3}$$

   This yields the negative gradient vector

$$\begin{aligned} U_k^{[m-1]} &= \left(U_{i,k}^{[m-1]}\right)_{i=1,\ldots,n} \\ &:= \left(-\frac{\partial}{\partial f_k} \rho\left(Y_i, \hat{f}^{[m-1]}(X_i), \hat{\sigma}^{[m-1]}\right)\right)_{i=1,\ldots,n} . \end{aligned} \tag{4}$$

   (c) Fit the negative gradient vector $U_k^{[m-1]}$ to each of the p components of $\boldsymbol{X}$ (i.e., to each predictor variable) separately by using p times the base-learner (regression estimator) specified in step 2. This yields p vectors of predicted values, where each vector is an estimate of the negative gradient vector $U_k^{[m-1]}$.

   (d) Select the component of $\boldsymbol{X}$ which fits $U_k^{[m-1]}$ best according to a pre-specified goodness-of-fit criterion. Set $\hat{U}_k^{[m-1]}$ equal to the fitted values from the corresponding best model fitted in 4(c).

   (e) Update $\hat{f}_k^{[m-1]} \leftarrow \hat{f}_k^{[m-1]} + \nu \hat{U}_k^{[m-1]}$, where $0 < \nu \leq 1$ is a real-valued step length factor.

   (f) For $k = 2, \ldots, K$ repeat steps 4(b) to 4(e). Update $\hat{f}^{[m]} \leftarrow \hat{f}^{[m-1]}$.

5. (a) Set $l = 0$.

   (b) Increase l by 1.

(c) Plug $\hat{f}^{[m]}$ and $\hat{\sigma}_1^{[m-1]}, \ldots, \hat{\sigma}_{l-1}^{[m-1]}, \hat{\sigma}_{l+1}^{[m-1]}, \ldots, \hat{\sigma}_L^{[m-1]}$ into the empirical risk function $\sum_{i=1}^n \rho(Y_i, f, \sigma)$ and minimize the empirical risk over σ_l. Set $\hat{\sigma}_l^{[m-1]}$ equal to the newly obtained estimate of σ_l.

(d) For $l = 2, \ldots, L$ repeat steps 5(b) and 5(c). Update $\hat{\sigma}^{[m]} \leftarrow \hat{\sigma}^{[m-1]}$.

6. Iterate Steps 3 to 5 until $m = m_{\text{stop}}$ for some stopping iteration m_{stop}.

The above algorithm can be viewed as a combination of two classical techniques for statistical model estimation, namely functional gradient descent (Friedman 2001) and backfitting (Hastie and Tibshirani 1990). In fact, each component f_k, $k = 1, \ldots, K$, is updated by

1. using the current estimates of the other components $f_1^*, \ldots, f_{k-1}^*, f_{k+1}^*, \ldots, f_K^*$ and $\sigma_1^*, \ldots, \sigma_L^*$ as offset values (*backfitting approach*, see step 4(b)), and by
2. adding an estimate of the true negative partial derivative $U_{i,k}^{[m-1]}$ to the current estimate of f_k^* (*gradient descent approach*, see step 4(e)).

After having obtained an update of $\hat{f}$ in step 4, the algorithm cycles through the components of σ, where in each step of the cycle, an update of σ_l, $l = 1, \ldots, L$, is obtained (step 5). This is accomplished by minimizing the empirical risk (evaluated at the current estimates of the other parameters f^* and $\sigma_1^*, \ldots, \sigma_{l-1}^*, \sigma_{l+1}^*, \ldots, \sigma_L^*$) numerically. A summary of the algorithm is given in Figure 1.

The value of the stopping iteration m_{stop} is the main tuning parameter of the algorithm. In the literature it has been argued that boosting algorithms should generally not be run until convergence. Otherwise, overfits resulting in suboptimal prediction rules are likely (see Bühlmann and Hothorn 2007). In this paper, five-fold cross-validation will be used for determining the value of m_{stop} (i.e., m_{stop} is the iteration with lowest predictive risk). The choice of the step length factor ν has been shown to be of minor importance with respect to the predictive performance of a boosting algorithm. The only requirement is that the value of ν is small ($0 < \nu \leq 0.1$), such that a stagewise adaption of the true prediction function is possible (see Bühlmann and Hothorn 2007, Schmid and Hothorn 2008a). In the following, a constant value of ν ($= 0.1$) will be used. In step 4(d) of the algorithm we will use the R^2 measure of explained variation as the goodness-of-fit criterion (since the vectors $U_k^{[m-1]}$ are measured on a continuous scale).

As outlined in Section 1, the algorithm combines model estimation with the selection of the most relevant predictor variables. In steps 4(c) to 4(e), by using a regression estimator as the base-learner, a structural relationship between $\boldsymbol{Y}$ and the set of predictors $\boldsymbol{X}$ is established. Due to the additive structure of the update (step 4(e)), the final estimates of $f_1^*, \ldots, f_K^*$ at iteration m_{stop} are fits of an additive model but will depend only on a subset of the p components of $\boldsymbol{X}$. In each iteration, the algorithm selects the basis direction "closest" to the descent direction of the prediction function (step 4(d)). Since only one element of $\boldsymbol{X}$ is used for updating the prediction function in step 4(e), the algorithm is applicable even if $p > n$. In this context, the proposed algorithm can be interpreted as a "stagewise regression" technique, as given by

Initialize: $\hat{f}_1^{[0]}, \ldots, \hat{f}_K^{[0]}$ and $\hat{\sigma}_1^{[0]}, \ldots, \hat{\sigma}_K^{[0]}$ with offset values.
  for $k = 1$ to K **do**
    Specify a base-learner for component f_k.
  end for
Evaluate:
  for $m = 1$ to m_{stop} **do**
    for $k = 1$ to K **do**
      (i) Compute $-\frac{\partial \rho}{\partial f_k}$ and evaluate at $\hat{f}^{[m-1]}(X_i), \hat{\sigma}^{[m-1]}, i = 1, \ldots, n$. This yields $U_k^{[m-1]}$.
      (ii) Fit $U_k^{[m-1]}$ to each of the p components of $\boldsymbol{X}$ separately by using p times the base-learner.
      (iii) Select the component of $\boldsymbol{X}$ which fits $U_k^{[m-1]}$ best. Set $\hat{U}_k^{[m-1]}$ equal to the fitted values from the best-fitting model.
      (iv) Update $\hat{f}_k^{[m-1]} \leftarrow \hat{f}_k^{[m-1]} + \nu \hat{U}_k^{[m-1]}$.
    end for
    Update $\hat{f}^{[m]} \leftarrow \hat{f}^{[m-1]}$.
    for $l = 1$ to L **do**
      Plug $\hat{f}^{[m]}$ and $\hat{\sigma}_1^{[m-1]}, \ldots, \hat{\sigma}_{l-1}^{[m-1]}, \hat{\sigma}_{l+1}^{[m-1]}, \ldots, \hat{\sigma}_L^{[m-1]}$ into the empirical risk function and minimize over σ_l. Set $\hat{\sigma}_l^{[m-1]}$ equal to the newly obtained estimate of σ_l.
    end for
    Update $\hat{\sigma}^{[m]} \leftarrow \hat{\sigma}^{[m-1]}$.
  end for

Figure 1: Gradient boosting with multidimensional prediction functions.

Efron et al. (2004). Note that the step length factor ν can be viewed as a regularization factor used for shrinking the predictions $\hat{f}^{[m]}$.

It is easily seen that in case of a one-dimensional prediction function $f^* \equiv f_1^*$ and an empty set of scale parameters, the boosting algorithm presented above reduces to the classical gradient descent algorithm developed by Friedman (2001). Similarly, if $f^* \equiv f_1^*$ and σ^* is one-dimensional, the algorithm is a generalization of the model fitting approach developed by Schmid and Hothorn (2008b) (where boosting was used for deriving parametric survival prediction rules). In case of a multinomial logit model with K outcome categories, the conditional probability of falling into category k is typically modeled via

$$\mathrm{P}\left(\boldsymbol{Y} = k | \boldsymbol{X}\right) = \frac{e^{f_k^*(\boldsymbol{X})}}{\sum_{j=1}^{K} e^{f_j^*(\boldsymbol{X})}} \ . \tag{5}$$

Thus, by setting σ^* equal to the empty set, the boosting algorithm introduced above can be used for fitting the multiclass model defined by (5). If the negative multinomial log likelihood

$$\begin{aligned} &\rho_{\text{multinom}}(Y, f) = \\ &-\sum_{i=1}^{n} \left[\log\left(n!\right) - \sum_{k=1}^{K} \mathrm{I}\left(Y_i = k\right) f_k(X_i) + \log\left(\sum_{j=1}^{K} e^{f_j(X_i)} \right) \right] \end{aligned} \tag{6}$$

is used as the loss function, the new algorithm will give essentially the same result as the multiclass method suggested by Friedman et al. (2000) and Hastie et al. (2003).

Finally, the boosting algorithm presented above can easily be modified such that the components of f are restricted to depend on subsets $\chi_1, \ldots, \chi_K \subset \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p\}$ only. Reducing the predictor spaces in step 4 of the algorithm adds considerable flexibility to the boosting procedure, since it allows for taking into account prior knowledge about the dependencies between f_k^* and $\boldsymbol{Y}$. A detailed example of this approach is given in Section 3.2.

3 Examples

3.1 Modeling nevus counts of preschool children

Nevus counts of children have been established as an important risk factor for malignant melanoma occurring later in life (Gallagher et al. 1990). In 1999 and 2000, the CMONDE Study Group (Uter et al. 2004, Pfahlberg et al. 2004) conducted a standardized skin assessment of $n = 3527$ preschool children in the German town of Göttingen. Nevus counts were collected in the course of a mandatory medical examination prior to school enrollment. Predictor variables in the data set included three continuous predictors (age, skin pigmentation, body mass index) and five categorical predictors (sex, hair color, skin type, color of iris, degree of freckling). The number of possible combinations of the categories was equal to 576.

In the following we will use the eight predictor variables for modeling expected nevus counts of children. In order to construct accurate predictions of the nevus counts, identification of relevant covariates is necessary. Also, given the fact that a relatively large number of categories is involved in the modeling process, some sort of regularization (i.e., shrinkage) of the prediction function is desirable. As pointed out earlier, the algorithm introduced in Section 2 is an appropriate technique for addressing these issues.

We will compare the predictions obtained from boosting with four different loss functions, where each loss function corresponds to a particular type of count data model:

1. *Negative Poisson log likelihood loss.* The most popular distribution used for modeling count data is the Poisson distribution. In the generalized linear model (GLM) setting (McCullagh and Nelder 1989), Poisson model estimates are obtained by maximizing the conditional log likelihood

$$l_{\text{Po}}(Y, f_1) = \sum_{i=1}^{n} \Big(Y_i \log\left(f_1(X_i)\right) - \log\left(Y_i!\right) - f_1(X_i) \Big) \tag{7}$$

over a one-dimensional prediction function $f = f_1$ (where $\exp(f_1(\boldsymbol{X}))$ corresponds to the conditional expectation of the outcome variable $\boldsymbol{Y}$ given the predictors $\boldsymbol{X}$). Since maximum likelihood estimation tends to become unstable in the presence of a larger

number of categorical predictors, we use the boosting algorithm introduced in Section 2 for obtaining estimates of the optimal prediction function f^*. This is accomplished by setting the loss function ρ equal to the *negative* Poisson log likelihood (7) and the set of scale parameters σ equal to the empty set.

2. *Negative NB log likelihood loss.* The underlying assumptions of the Poisson model are often too restrictive for capturing the full variability contained in a data set. A common way to model over-dispersed data is to consider negative binomial (NB) regression models. The log likelihood corresponding to the negative binomial model is given by

$$\begin{aligned} l_{\mathrm{NB}}(Y, f_1, \sigma_1) &= \sum_{i=1}^{n} \Big(\log\left[\Gamma(Y_i + \sigma_1)\right] - \log\left(Y_i!\right) - \log\left[\Gamma(\sigma_1)\right] \Big) \\ &+ \sum_{i=1}^{n} \sigma_1 \log\left(\frac{\sigma_1}{f_1(X_i) + \sigma_1} \right) \\ &+ \sum_{i=1}^{n} Y_i \log\left(\frac{f_1(X_i)}{f_1(X_i) + \sigma_1} \right) , \end{aligned} \tag{8}$$

where $f = f_1$ is a one-dimensional prediction function and $\sigma = \sigma_1$ is a one-dimensional scale parameter used for modeling the variance of $\boldsymbol{Y}$. It is well known that $\lambda := \exp(f_1(\boldsymbol{X}))$ corresponds to the conditional expectation of the outcome variable $\boldsymbol{Y}$ given the predictors $\boldsymbol{X}$, and that the conditional variance of $\boldsymbol{Y}|\boldsymbol{X}$ is given by $\lambda + \lambda^2/\sigma_1$. The log likelihood given in (8) reduces to a Poisson log likelihood as $\sigma_1 \to \infty$. In the following we will use the boosting algorithm introduced in Section 2 for obtaining estimates of the optimal parameters f_1^* and σ_1^*. This is accomplished by setting ρ equal to the negative NB log likelihood (8) and σ equal to the scale parameter σ_1 in (8).

3. *Negative zero-inflated Poisson log likelihood loss.* Excessive amounts of zero counts, i.e., more zeros than expected in a Poisson or negative binomial model, are a common problem associated with count data. In case of the CMONDE data, the fraction of zero nevus counts is approximately 8.4%, which is about 25 times as much as the corresponding fraction to be expected from the unconditional distribution of the nevus counts (0.337%). In order to take this problem into account, we additionally fit a zero-inflated Poisson model to the CMONDE data. The log likelihood of the zero-inflated Poisson model is given by

$$\begin{aligned} l_{\mathrm{ZIPo}}(Y, f_1, f_2) &= \sum_{i:Y_i=0} \left(\frac{e^{f_1(X_i)}}{1 + e^{f_1(X_i)}} + \frac{1}{1 + e^{f_1(X_i)}} e^{-e^{f_2(X_i)}} \right) \\ &+ \sum_{i:Y_i>0} \log\left(\frac{1}{1 + e^{f_1(X_i)}} \right) \\ &+ \sum_{i:Y_i>0} \left(Y_i \cdot f_2(X_i) - \log\left(Y_i!\right) - e^{f_2(X_i)} \right) , \end{aligned} \tag{9}$$

where f_1 is the predictor of the binomial logit model

$$\mathrm{P}\left(\boldsymbol{Z}=0|\boldsymbol{X}\right)=\frac{e^{f_1(\boldsymbol{X})}}{1+e^{f_1(\boldsymbol{X})}} \tag{10}$$

with a binary outcome variable $\boldsymbol{Z}\in\{0,1\}$, and f_2 is the predictor of the Poisson model

$$\mathrm{P}\big(\tilde{\boldsymbol{Y}}=k|\boldsymbol{X}\big)=\frac{e^{k\cdot f_2(\boldsymbol{X})}}{k!}\,e^{e^{f_2(\boldsymbol{X})}} \tag{11}$$

with a Poisson-distributed outcome variable $\tilde{\boldsymbol{Y}}$. It is easily seen from (9) to (11) that the zero-inflated Poisson model is a mixture of a point mass at zero (accounting for an extra amount of zeros) and a Poisson distribution. For details we refer to Hilbe (2007). Since we want to regularize the estimates of both components of the prediction function, we use the boosting algorithm introduced in Section 2. This is achieved by setting ρ equal to the negative log likelihood (9) and $f=(f_1,f_2)$. The set of scale parameters σ is set equal to the empty set.

4. *Negative zero-inflated NB log likelihood loss.* In case of over-dispersed data, modeling additional amounts of zero counts can be accomplished by using the zero-inflated negative binomial model. The log likelihood of this model is given by

$$\begin{aligned} l_{\mathrm{ZINB}}(Y,f_1,f_2,\sigma_1) \;=\; & -\sum_{i=1}^{n}\log\left(1+e^{f_1(X_i)}\right) \\ & +\sum_{i:Y_i=0}\log\left(e^{f_1(X_i)}+\left(\frac{e^{f_2(X_i)}+\sigma_1}{\sigma_1}\right)^{-\sigma_1}\right) \\ & -\sum_{i:Y_i>0}\sigma_1\log\left(\frac{e^{f_2(X_i)}+\sigma_1}{\sigma_1}\right) \\ & +\sum_{i:Y_i>0}Y_i\log\left(1+e^{-f_2(X_i)}\cdot\sigma_1\right) \\ & -\sum_{i:Y_i>0}\left(\log\left(\Gamma(\sigma_1)\right)+\log\left(1+Y_i\right)\right) \\ & -\sum_{i:Y_i>0}\log\left(\sigma_1+Y_i\right)\;. \end{aligned} \tag{12}$$

Similar to the zero-inflated Poisson model, the zero-inflated negative binomial model is a mixture of a point mass at zero (modeled by a binomial GLM) and a zero-inflated negative binomial regression model. We apply the new boosting algorithm to the CMONDE data by setting ρ equal to the negative log likelihood (12) and $f=(f_1,f_2)$. The set of scale parameters σ will be set equal to the scale parameter σ_1 in (12).

In order to compare the four models described above, we carried out a benchmark study using the CMONDE data. In a first step, the full data set was randomly split into 20 pairs of training

samples and test samples. Each training sample contained 3175 observations, i.e., about 90% of the data. In a next step, the boosting algorithm introduced in Section 2 was used to estimate the parameters of the four count data models. As base-learners, simple *linear* regression models were used, so that the components of $\hat{f}$ are linear functions of the predictors. As a consequence of this strategy, coefficient estimates were obtained for each predictor variable. For all components of f, variables were selected from the full set of predictors (i.e., no restrictions were made to the set of predictors at the beginning of the algorithm). In a last step, the prediction rules obtained from the four models were evaluated by using the 20 test samples. All computations were carried out with the R system for statistical computing (version 2.7.2, R Development Core Team 2008) using a modification of the `glmboost()` function in package **mboost** (version 1.0-4, Hothorn et al. 2008). In order to determine the 20 values of the stopping iteration m_{stop}, we ran five-fold cross-validation on the training samples.

Since the negative versions of the log likelihood functions (7), (8), (9) and (12) are used as loss functions for the respective boosting algorithms, it would be a natural approach to measure the prediction accuracy of the boosting methods by computing the predictive log likelihood values from the test samples. Since the functions (7), (8), (9) and (12) are measured on different scales, however, using this approach would be unsuitable for comparing the four models. We therefore used the Brier score (Brier 1950), which is a model-independent measure of prediction accuracy. The Brier score is defined as the negative average squared distance between the observed probabilities and the predicted probabilities of the outcome categories in the test samples. Thus, since the count data models under consideration are characterized by equidistant outcome categories, using the Brier score corresponds to using the integrated squared difference of the estimated and observed c.d.f.'s of the test observations as a measure of prediction error.

More formally, the Brier score for test sample t, $t \in \{1, \ldots, 20\}$, is defined as

$$BS_t := -\frac{1}{n_t} \sum_{k=1}^{M} \sum_{i=1}^{n_t} \left(p_{ikt} - \hat{p}_{ikt}\right)^2 \ , \tag{13}$$

where n_t is the number of observations in test sample t and M is the number of categories of the outcome variable. Let Y_{it} be the i-th realization of the outcome variable $\boldsymbol{Y}$ in test sample t. Then the parameters p_{ikt} in (13) are defined as

$$p_{ikt} = \begin{cases} 1 & \text{if } Y_{it} = k \\ 0 & \text{otherwise} \end{cases} \ . \tag{14}$$

The parameters p_{ikt} can be interpreted as the observed probability of category k given X_{it} (where X_{it} denotes the i-th realization of the predictor variable $\boldsymbol{X}$ in test sample t). Similarly, the predicted probabilities of category k given X_{it} (denoted by $\hat{p}_{ikt}$) are obtained by plugging the estimates of f^* and σ^* (computed from *training* sample t) into the likelihood functions corresponding to *test* sample t. For computational reasons we define $p_{iMt} := 1 - \sum_{k \leq M^*} p_{ikt}$, where M^* is the largest outcome value observed in the data.

The Brier score can generally be used for assessing the quality of probabilistic forecasts. It is an example of a so-called "proper" scoring rule, where "proper" means that the expectation

of (13) is maximized if the predictions $\hat{p}_{ikt}$ are computed from the true model with parameters f^* and σ^*. For details on proper scoring rules we refer to Gneiting and Raftery (2007). Generally, a large value of the Brier score corresponds to a highly accurate prediction rule (and vice versa).

In Figure 2, boxplots of the Brier score values computed from the 20 test samples of the CMONDE data are shown. Obviously, the Brier score values corresponding to the Poisson model are smallest on average, indicating that not all of the variability contained in the data is captured by the Poisson distribution. It can also be seen from Figure 2 that introducing a scale parameter for modeling overdispersion, i.e., using a negative binomial model, leads to improved predictions. On the other hand, the prediction accuracy of the negative binomial model is substantially higher than the accuracy of the zero-inflated Poisson model. Similarly, if compared to the negative binomial model, the zero-inflated negative binomial model does not seem to lead to an additional increase in predictive power. This result is confirmed by the Vuong test ($p = 0.112$), suggesting that there is no significant difference between the negative binomial model and its zero-inflated extension (cf. Greene 1994). For this reason, and also because of its simpler structure, we suggest to use the negative binomial model for predicting nevus counts. (Note, however, that the Vuong statistic cannot be interpreted strictly in this context, as the estimates of f^* and σ^* are not the maximum likelihood estimates.)

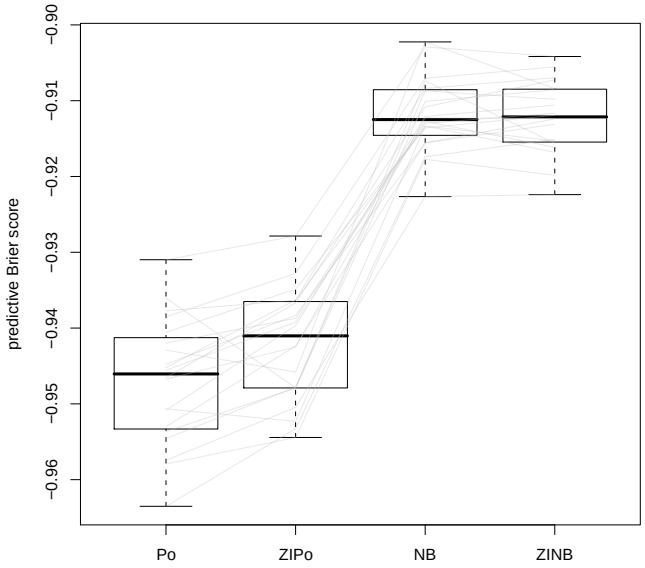


Figure 2: Predictive Brier score values computed from the 20 test samples of the CMONDE data (Po = Poisson model, ZIPo = zero-inflated Poisson model, NB = negative binomial model, ZINB = zero inflated negative binomial model). Obviously, the negative binomial models perform better than the Poisson models, indicating that overdispersion is present in the CMONDE data.

The coefficient estimates of the negative binomial model (computed from the full data set) are shown in Table 1. When comparing the directions of the estimates to the results published by the CMONDE study group, it becomes obvious that the original results (Pfahlberg et al. 2004) are supported by the boosting estimates. For example, children with blonde hair tend to have substantially more nevi than children with black hair. Also, the number of facial freckles is positively correlated with the number of nevi, where male children have significantly more nevi (on average) than female children. A detailed description of the variables contained in Table 1 and their effect on nevus counts can be found in Pfahlberg et al. (2004).

Predictor	Est. Coefficient	95% CI
Intercept	-0.14068	[-0.51618 , -0.07116]
sex (male)	0.00000	
sex (female)	-0.14582	[-0.17535 , -0.06107]
hair color (blonde)	0.00000	
hair color (brown)	0.01013	[-0.04855 , 0.04980]
hair color (red)	-0.46247	[-1.11221 , -0.40183]
hair color (black)	-0.71926	[-0.81337 , -0.48059]
Fitzpatrick skin type (I)	0.00000	
Fitzpatrick skin type (II)	0.23245	[0.11955 , 0.44403]
Fitzpatrick skin type (III)	0.16916	[0.06052 , 0.36555]
Fitzpatrick skin type (IV)	0.05135	[-0.05836 , 0.26312]
color of iris (blue)	0.00000	
color of iris (dark brown)	-0.40879	[-0.50766 , -0.27041]
color of iris (green-blue)	-0.14020	[-0.21713 , -0.01745]
color of iris (green-brown)	-0.09826	[-0.19544 , -0.03328]
color of iris (light blue)	-0.01592	[-0.14156 , 0.08361]
color of iris (light brown)	-0.22140	[-0.30272 , -0.15544]
facial freckles (none)	0.00000	
facial freckles (few)	0.14717	[0.08570 , 0.20098]
facial freckles (many)	0.27431	[0.13412 , 0.49588]
skin pigmentation		
(reflectance in percent at 650 nm)	0.01524	[0.00493 , 0.03437]
age in years	0.11097	[0.02410 , 0.19061]
body mass index in kg/m^2	0.03901	[0.01759 , 0.05234]
$\hat{\sigma}$	1.94294	[1.76163 , 1.98270]

Table 1: Boosting coefficient estimates obtained from the CMONDE data (negative binomial model). The 95% confidence intervals were computed from 50 bootstrap samples. Skin type was determined by using the categories proposed by Fitzpatrick (Fitzpatrick 1988). Skin pigmentation was quantified with remission photometry, i.e., small reflectance measurements correspond to a highly pigmented skin.

3.2 Breast cancer gene expression data

In this section we will re-analyze a breast cancer data set collected by the Netherlands Cancer Institute (van't Veer et al. 2002, van de Vijver et al. 2002). Based on $n = 295$ patients and the expressions of 70 genes, van't Veer et al. (2002) proposed a signature for prediction of metastasis-free survival in breast cancer. In addition to gene expression measurements, the data contained nine clinical covariates (such as age and tumor diameter), which are well-established predictors of metastasis-free survival. It is therefore desirable to construct a combined predictor that makes use of both the gene expression data and the clinical covariates. A particularly important issue in this context is that using clinical covariates for prediction is often mandatory for practical and conceptual reasons. For reasons of interpretability, sparse models are desired, so that only a small number of genes should contribute to the combined predictor. We therefore want to construct a predictor depending on a small number of genes but with a mandatory inclusion of all clinical covariates.

In the following we will use the boosting algorithm introduced in Section 2 to address this problem. We define the prediction function f^* as follows:

$$f^* = (f_1^*, \ldots, f_K^*) := f_1^* + \cdots + f_K^* \ . \tag{15}$$

The components $f_1^*, \ldots, f_{K-1}^*$ in (15) correspond to the effects of the clinical covariates, while the component f_K^* corresponds to the effect of the gene expression measurements. In case of the breast cancer data, there are nine clinical covariates $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_9$, i.e., f^* depends on $K = 10$ components. Having defined the prediction function, we restrict the components $f_1^*, \ldots, f_9^*$ to depend on the "single covariate" sets $\chi_1 := \{\boldsymbol{X}_1\}, \ldots, \chi_9 := \{\boldsymbol{X}_9\}$, respectively. Similarly, the component f_{10}^* is restricted to depend on the set $\chi_{10} := \{\boldsymbol{X}_{10}, \ldots, \boldsymbol{X}_{79}\}$ corresponding to the 70 gene expression measurements proposed by van't Veer et al. (2002). In other words, the definition of the sets $\chi_1, \ldots, \chi_9$ leads to mandatory updates of the clinical components $f_1^*, \ldots, f_9^*$ in each iteration of the boosting algorithm. In contrast, the component f_{10}^* corresponding to the gene expression data is updated by selecting only one gene in each iteration. As a consequence, the model fit will depend on all clinical covariates but only on a small number of genes.

As a loss function we use the negative partial log likelihood function of a Cox proportional hazards model. Thus, if simple linear regression models are used as base-learners, the final boosting estimate $\hat{f}$ can be interpreted as the linear predictor of the a Cox model (Ridgeway 1999). Note that fitting Cox models to gene expression data is a well-established approach, which has been applied to the breast cancer data before (see van Houwelingen et al. 2006).

In order to compare the predictive accuracy of the combined predictor with other types of predictors, we carried out a benchmark study using the breast cancer data. In a first step, the full data set was randomly split into 20 pairs of training samples and test samples. In contrast to the CMONDE data discussed in Section 3.1, we decreased the size of the training sets to two thirds of the data (197 observations). This strategy lead to an increased number of observations in the test samples, thus assuring that reliable estimates of the prediction error could be computed. In a next step, three estimation techniques were applied to the training samples:

(a) boosting with a mandatory inclusion of the clinical predictors (as described above),

(b) boosting with the set of gene expression measurements χ_{10} only,

(c) a Cox proportional hazards model with the clinical predictors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_9$ only.

In other words, we compared the combined predictor to the predictor based on the gene expression data only (method (b)) and the predictor based on the clinical data only (method (c)). As base-learners, simple linear regression models were used.

In order to evaluate the prediction rules obtained from the training samples, we used the predictive partial log likelihood values computed from the test samples. Using this approach for comparing the three estimation techniques is an appropriate strategy, since the same loss function (namely the negative partial log likelihood function) is used for all three techniques. Therefore, for each technique, the predictive partial log likelihood values are measured on the same scale.

Figure 3 shows boxplots of the average partial log likelihood values obtained from the 20 test samples of the breast cancer data. Both the combined predictor and the predictor based on the gene expression data seem to perform better than predictor based on the clinical data only. On the other hand, the average difference between the predictions obtained from the

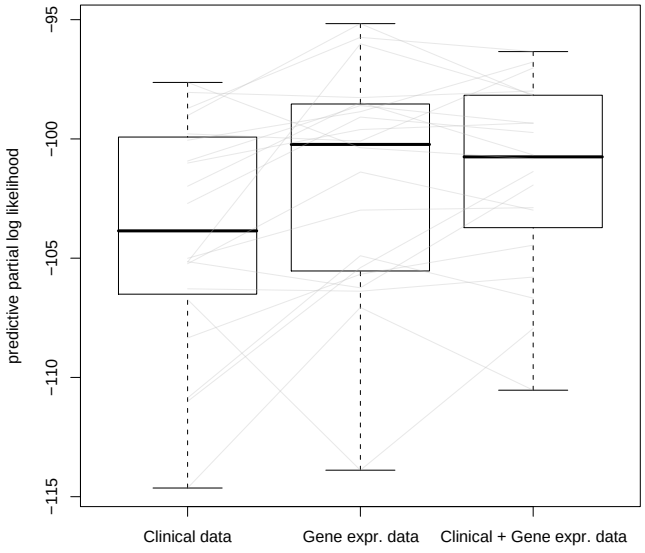


Figure 3: Average predictive partial log likelihood values computed from the 20 test samples of the breast cancer data. The prediction accuracy of the combined predictor is similar to that of the gene-based predictor. Both the combined predictor and the gene-based predictor result in more accurate predictions than the predictor using the clinical data only.

combined predictor and the predictions obtained from the gene-based predictor seems to be small. However, the variance of the predictive partial log likelihood values obtained from the combined predictor is somewhat smaller than that of the gene-based predictor. These results suggest that in case of the breast cancer data, using gene expression measurements leads to more accurate predictions than using clinical data only. However, the predictive accuracy of the gene-based predictor can only be slightly improved if a combination of the clinical data and the gene expression data is used.

This conclusion is further confirmed if, for the full data set, the values of the estimated prediction function $\hat{f}(X_i)$ are split into a "low-risk" group, a "medium-risk" group and a "high-risk" group. Kaplan-Meier estimates of the survival curves corresponding to the three groups are shown in Figure 4. Obviously, the combined predictor performs better than the clinical predictor with respect to separating the three risk groups. On the other hand, the differences between the Kaplan-Meier estimates corresponding to the combined predictor and the gene-based predictor are small.

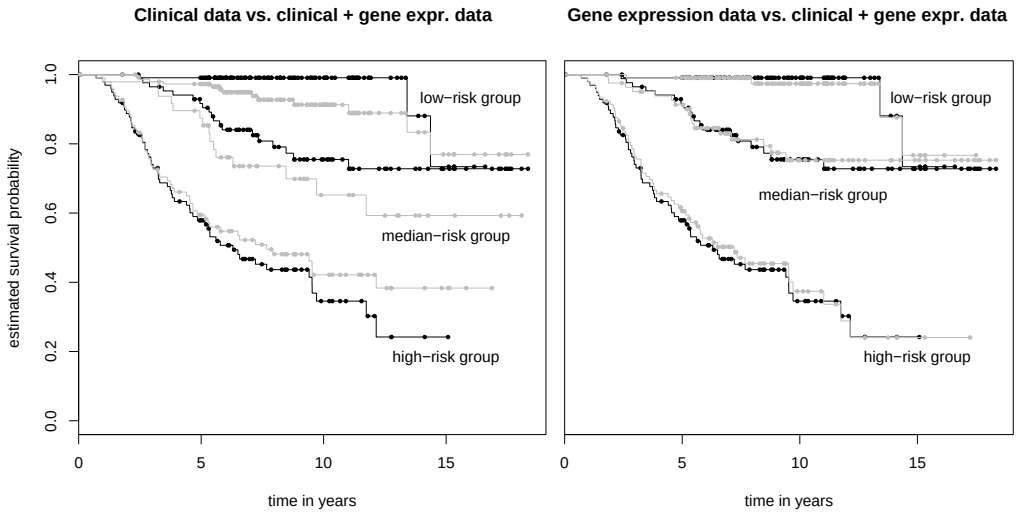


Figure 4: Kaplan-Meier estimates of the three risk groups, as computed from the full breast cancer data set. Conditional inference survival trees were used to determine the split points for the estimated prediction function $\hat{f}$ (see Hothorn et al. 2006). In the left panel, the predictor based on the clinical data (grey lines) is compared to the predictor based on both the clinical data and the gene expression data (black lines). Obviously, the combined predictor improves the separation of groups with different mortality levels, at least within a time range of $[0, 10]$ years. In the right panel, the predictor based on the gene expression data only (grey lines) is compared to the predictor based on both the clinical data and the gene expression data (black lines). Obviously, if compared to the gene-based predictor, the combined predictor only leads to a slightly improved separation of groups with different mortality levels.

4 Conclusion

Originally developed as a machine learning technique for predicting binary outcomes (Freund and Schapire 1997), boosting has gained considerable attention in the statistical community over that last years. Most notably, by showing that the original boosting algorithm for binary classification can be interpreted as a gradient descent technique for minimizing arbitrary loss functions, Breiman (1998, 1999) has laid the foundations for applying boosting algorithms to a wide class of statistical estimation problems. Later, by introducing the "statistical view" of boosting, Friedman et al. (2000) have established boosting as a tool for fitting general types of regression models.

In this paper we have extended the gradient descent approach of Friedman (2001) by constructing a boosting algorithm for regression models with multidimensional prediction functions. Instead of descending the gradient in one direction only, the algorithm introduced in this paper successively computes the *partial* derivatives of the components of the prediction function. Updates of a component of the prediction function are then computed by using the current values of the other components as offset values. The new algorithm can therefore be viewed as a combination of the original gradient boosting approach and the backfitting idea introduced by Hastie and Tibshirani (1990). Most important, the regularization concept of the original boosting approach carries over to the new multi-dimensional algorithm. As a result, the algorithm introduced in this paper is particularly useful for analyzing high-dimensional data (where selecting a moderate number of relevant predictors is often a key problem).

As demonstrated in Section 3, the algorithm proved to work well in two important statistical modeling situations, namely count data models and survival predictions based on gene expression data. Concerning count data models, we found that boosting is a suitable technique for fitting negative binomial models and zero-inflated count data models (involving a two-dimensional prediction function and the estimation of a dispersion parameter). Apart from the models considered in this paper, the algorithm can easily be used to fit other popular types of count data models, such as the Hurdle model (Mullahy 1986) or the generalized Poisson distribution (Consul and Jain 1973). Fitting these models to the CMONDE data did not lead to an improvement in predictive accuracy (if compared to the negative binomial model).

The problem of including mandatory clinical covariates into survival predictions based on gene expression data has first been investigated by Binder and Schumacher (2008). Instead of updating each clinical covariate separately, the authors constructed a boosting algorithm which, in each iteration, computes a simultaneous update of all clinical covariates. Afterwards, the best fitting predictor is selected out of the set of gene expression data in each iteration. Instead of shrinking the estimates (with a step length factor $\nu < 1$), Binder and Schumacher (2008) used one-step Fisher scoring and a penalized base-learning procedure for the gene expression data. An alternative strategy, which is similar to the one used in this paper, has been suggested by Bühlmann and Hothorn (2007), Section 7.3. With this strategy, the prediction function corresponding to the clinical covariates is first estimated by using a classical regression model. Afterwards, the predicted values of this model are used as offset values for a gradient boost-

ing algorithm with the gene expression data only. A disadvantage of this two-step approach is that the multivariate structure between the clinical covariates and the genes may not be fully explored. In contrast, the method presented in this paper uses an a priori restriction of the subsets of predictor variables, thus constituting a natural example of the more general boosting framework introduced in Section 2 of this paper. Furthermore, by using a shrinkage factor ν instead of working with Fisher scoring and penalized base-learners, the algorithm presented in this paper is more similar to the classical forward stagewise idea (Efron et al. 2004) than the algorithm developed by Binder and Schumacher (2008).

A different approach to modeling the breast cancer data would be to use Cox models with a ridge penalty, as suggested by van Houwelingen et al. (2006). This approach, however, does not allow for selecting the most relevant predictor variables but instead results in (shrunken) coefficient estimates for all 70 genes. Since a major focus of this paper is on variable selection (leading to sparse model fits with a good interpretability), we did not consider ridge regression methods.

In addition to the application examples presented in Section 3, the proposed algorithm is generally suitable for solving a wide class of estimation problems with multidimensional prediction functions. In particular, boosting constitutes a natural approach to estimating the parameters of (identifiable) finite mixture models, where, in addition to the regression parameters, the class probabilities of a fixed number of latent categories have to be estimated. (Note that the zero-inflated count data models considered in Section 3.1 are special cases of finite mixture models.) Furthermore, the algorithm can easily be modified such that different types of base-learners can be applied to different components of the prediction function. For example, one could use smooth base-learners for modeling the first component of the prediction function, tree base-learners for modeling the second component, linear base-learners for modeling the third component, etc. Similarly, by using two-dimensional base-learners, interaction terms between the covariates can be included into the prediction function. In addition, instead of updating scale parameters by numerical optimization, it is possible to regress them on a (possibly restricted) set of covariates. This can easily be accomplished by treating the (sub)set of scale parameters as an ordinary component of the prediction function, i.e., by modeling the scale parameter(s) via the same base-learning procedures as used for the prediction function. In this regard, boosting with multidimensional prediction functions constitutes a highly flexible approach to statistical model estimation.

Acknowledgements

The authors thank the Regional Health Authority of the city of Göttingen (Head: Dr. W. R. Wienecke) and Prof. Dr. K. F. Kölmel (Department of Dermatology at the University of Göttingen) for the permission to use the data of the CMONDE study for illustrative purposes. The CMONDE study was supported by the Cancer Fund of Lower Saxony. MS and TH were supported by Deutsche Forschungsgemeinschaft (DFG), grant HO 3242/1–3. SP was supported by DFG, collaborate research area SFB 539-A4/C1.

References

Binder, H. and M. Schumacher (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics 9*.

Breiman, L. (1998). Arcing classifiers (with discussion). *The Annals of Statistics 26*, 801–849.

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

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3.

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

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science 22*, 477–522.

Bühlmann, P. and B. Yu (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–338.

Bullinger, L., K. Döhner, E. Bair, S. Fröhlich, R. F. Schlenk, R. Tibshirani, H. Döhner, and J. R. Pollack (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myloid leukemia. *New England Journal of Medicine 350*, 1605–1616.

Consul, P. and G. Jain (1973). A generalization of the poisson distribution. *Technometrics 15*, 791–799.

Dudoit, S., J. Fridlyand, and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association 97*, 77–87.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fitzpatrick, T. B. (1988). The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology 124*, 869–871.

Freund, Y. and R. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*, 119–139.

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

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

Gallagher, R. P., D. I. McLean, C. P. Yang, A. J. Coldman, H. K. Silver, J. J. Spinelli, and M. Beagrie (1990). Suntan, sunburn, and pigmentation factors and the frequency of acquired melanocytic nevi in children. similarities to melanoma: the Vancouver mole study. *Archives of Dermatology 126*, 770–776.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfeld, and E. S. Lander (1999).

Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. Working papers ec-94-10, Department of Economics, New York University, Leonard N. Stern School of Business.

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

Hastie, T., R. Tibshirani, and J. Friedman (2003). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). New York: Springer.

Hilbe, J. M. (2007). *Negative Binomial Regression*. Cambridge: Cambridge University Press.

Hothorn, T., P. Bühlmann, T. Kneib, and M. Schmid (2008). *mboost: Model-Based Boosting*. R package version 1.0-2. http://R-forge.R-project.org.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*, 651–674.

International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993). A predictive model for aggressive non-Hodgkin's lymphoma: The international non-Hodgkin's lymphoma prognostic factors project. *New England Journal of Medicine 329*, 987–994.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman & Hall.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics 33*, 341–365.

Park, M. Y. and T. Hastie (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B 69*, 659–677.

Pfahlberg, A., W. Uter, C. Kraus, W. R. Wienecke, U. Reulbach, K. F. Kölmel, and O. Gefeller (2004). Monitoring of nevus density in children as a method to detect shifts in melanoma risk in the population. *Preventive Medicine 38*, 382–387.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics 31*, 172–181.

Schmid, M. and T. Hothorn (2008a). Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis*. To appear.

Schmid, M. and T. Hothorn (2008b). Flexible boosting of accelerated failure time models. *BMC Bioinformatics 9:269*.

Segal, M. R. (2006). Microarraygene expression data with linked survival phenotypes: Diffuse large-B-cell lymphoma revisited. *Biostatistics 7*, 268–285.

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

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society, Series B 67*, 91–108.

Uter, W., A. Pfahlberg, B. Kalina, K. F. Kölmel, and O. Gefeller (2004). Inter-relation between variables determining constitutional UV sensitivity in Caucasian children. *Photodermatology, Photoimmunology & Photomedicine 20*, 9–13.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen,

A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine 347*, 1999–2009.

van Houwelingen, H. C., T. Bruinsma, A. A. M. Hart, L. J. van't Veer, and L. F. A. Wessels (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine 25*, 3201–3216.

van't Veer, L. J., H. Y. Dai, M. J. van de Vijver, Y. D. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernard, and S. H. Friend (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*, 530–536.

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

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

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