Rank-based p-values for sparse high-dimensional risk prediction models fitted by componentwise boosting

Harald Binder, Christine Porzelius & Martin Schumacher

Universität Freiburg i. Br.

FDM-Preprint Nr. 101 December 2009

Zentrum für Datenanalyse und Modellbildung

Universität Freiburg

Eckerstraße 1

D–79104 Freiburg im Breisgau

und

Institut für Medizinische Biometrie und Medizinische Informatik

Universitätsklinikum Freiburg

Stefan-Meier-Straße 26

D–79104 Freiburg im Breisgau

binderh@fdm.uni-freiburg.de

cp@fdm.uni-freiburg.de

ms@imbi.uni-freiburg.de

Abstract

There are many techniques for fitting sparse risk prediction models to high-dimensional data, e.g., allowing to predict a clinical endpoint by automatically determining a small set of informative covariates from gene expression measurements. Techniques for regularized estimation, such as componentwise boosting, provide sparse estimates for linear predictors. Many estimated parameters will be equal to zero, resulting in good prediction performance and a short list of informative genes. However, typically no p-values for single genes will be available, which is a downside compared to univariate test-based approaches. We address this by presenting an approach for obtaining p-values for componentwise likelihood-based boosting. The latter technique performs a large number of model fitting steps, where in each step one covariate is selected based on a score statistic. For quantifying uncertainty, we propose to consider the ranking based on the score statistic for all covariates in each step. Comparing the median rank of a covariate across boosting steps to a null distribution, obtained from data with permuted response, provides a p-value that takes the multivariable structure into account. In a simulation study, these rank-based p-values for multivariable models are seen to provide competitive performance concerning identification of informative covariates. In contrast to univariate p-values, the actual Type I error rates are close to the nominal level in most settings. In an application to time-to-event data with gene expression measurements from patients with diffuse large-B-cell lymphoma, the proposed p-values are seen to provide valuable complementary information to the fitted model. This combines the advantages of multivariable risk prediction techniques and univariate approaches for quantifying uncertainty.

Keywords: High-dimensional data; Risk prediction; Gene lists; Sparseness; Boosting; Uncertainty.

1 Introduction

Modern molecular measurement techniques, such as gene expression microarrays, often result in a large number of variables for a relatively small number of patients. Based on the specific objective, there are different strategies for dealing with such data. When comparison between groups is wanted, e.g., "cases" vs. "controls" or gene expression from different types of tissues, typically test-based techniques are employed. There, one test statistic is calculated for each molecular entity, e.g., for each gene, resulting in a list of p-values for quantifying uncertainty, e.g., with respect to the group difference for each single gene. When selecting a subset of genes to be called differentially expressed, the false discovery rate with respect to that list can be controlled (see Benjamini and Hochberg, 1995, for example).

However, when high-dimensional molecular measurements are to be linked to a future clinical endpoint, e.g., "dead vs. alive" or "time to relapse", techniques for fitting of risk prediction models may be more appropriate. Often, fitting of a regression model with simultaneous variable selection is wanted, resulting in sparse models, i.e., models that contain only a small number of automatically selected molecular covariates. For example, with gene expression data this promises a relatively short list of genes that is informative with respect to future clinical events.

There are many techniques that provide such sparse fits. For example, the lasso (Tibshirani, 1996) or componentwise boosting (Bühlmann and Yu, 2003) allow for sparse estimation of the parameter vector of standard regression models in a high-dimensional setting. These have been adapted to various kinds of endpoints (see Park and Hastie, 2007, for an efficient algorithm for the lasso with binary response and time-to-event data, and Tutz and Binder, 2007; Binder and Schumacher, 2008b, for componentwise boosting, for example).

As these techniques have been designed to provide risk prediction models, the resulting models are naturally evaluated with respect to prediction performance (see Schumacher et al., 2007, for example). However, if a model is found to provide good prediction performance, this only provides general information about the employed small number of molecular quantities. Prediction performance does not provide specific information concerning the contribution of a single gene to the model. In particular, the uncertainty with respect to this contribution remains unclear. In a low-dimensional setting, confidence intervals and p-values would provide such information. However, in a high-dimensional setting, variable selection often will defeat attempts to obtain reliable confidence intervals (see Leeb and Pötscher, 2006, for example). Also, there are hardly any techniques that provide sparse risk prediction models and p-values. On the other side, Bayesian techniques, which excel at quantifying uncertainty (see Noma et al., 2009, for example), mostly do not provide one single sparse risk prediction model (see Kaderali et al., 2006, for one such approach).

In the following, we present an approach for obtaining p-values for one specific class of sparse model fitting techniques, namely componentwise likelihood-based boosting (Tutz and Binder, 2007; Binder and Schumacher, 2008b). These techniques employ a stagewise approach for building up the estimated parameter vector of a sparse high-dimensional regression model. As will be shown, p-values can be obtained by employing information from this stagewise process by assigning ranks to each covariate in each step and comparing aggregated ranks to a permutation-based null distribution.

The primary aim of the proposed approach for obtaining p-values is to quantify uncertainty of regression model components fitted to high-dimensional data. Therefore, the focus will not be on exact distributional properties of the corresponding test statistic, but on practical performance in typical high-dimensional settings. Correspondingly, also other names, instead of "p-values", could be chosen, for indicating their nature as in-

dices of uncertainty, without a strong focus on theoretical foundations. However, as the proposed procedure for obtaining these indices of uncertainty closely mimics other statistical testing approaches, they will be called p-values in the following.

Several other approaches have been proposed for evaluating stability of sparse risk prediction models or of gene lists and rankings. For example, bootstrap inclusion frequencies are obtained by performing variable selection in bootstrap data sets, generated by sampling from the original data, and recording the proportion of bootstrap data sets where, e.g., a gene has been selected (see Qiu et al., 2006, for example). Similarly, Baker and Kramer (2006) consider selection frequencies for informative genes in combination with prediction performance for gene sets of various sizes. Baek et al. (2009) show that improved performance can be obtained by using frequently selected genes as a starting point for model building. Boulesteix and Slawski (2009) provide an overview of techniques for quantifying the stability of ranked gene lists. However, none of these approaches provides p-values for model components, which are are easily obtained for single genes from test-based techniques. Meinshausen et al. (2009) propose a technique using multiple random splits, where variable selection is performed in one data set and p-values are obtained from the remaining data. However, aggregation of p-values from the various splits is difficult, and p-values are provided only for selected genes. In contrast, the approach proposed in the following attempts to provide p-values for all covariates, when fitting models by componentwise boosting.

An overview of componentwise likelihood-based boosting will be given in Section 2.1. The proposed approach for extracting p-values then is described in Section 2.2. The performance of the p-values with respect to identification of informative covariates and their actual Type I error levels are evaluated in a simulation study in Section 3, specifically comparing the proposal to p-values from univariate approaches. An application example with gene expression data and a time-to-event endpoint for patients with diffuse

large-B-cell lymphoma is given in Section 4. Section 5 provides some concluding remarks and a discussion about the generality of the proposed technique for obtaining p-values for sparse high-dimensional risk prediction models.

2 Rank-based p-values for componentwise boosting

2.1 Componentwise likelihood-based boosting

When fitting generalized linear models (McCullagh and Nelder, 1989) by componentwise likelihood boosting (Tutz and Binder, 2007), there are observations $(x_i, y_i), i = 1, \ldots, n$, with potentially high-dimensional covariate vectors $x_i = (x_{i1}, \ldots, x_{ip})'$. The response y_i is taken to be from an exponential family, e.g., it may be continuous or binary, where the latter is typical for clinical endpoints that are to be linked to high-dimensional molecular data. The structural part of the model then is given by

$$E(y_i|x_i) = g(\eta_i) = g(\beta_0 + x_i'\beta),$$

where g is a known response function. The parameter vector $\beta = (\beta_1, \ldots, \beta_p)$ (and the intercept term β_0, which will be neglected in the following) is typically estimated by maximizing the log-likelihood $l(\beta)$ (for more details see McCullagh and Nelder, 1989). However, this is no longer possible when the number of covariates p is larger than the number of observations n.

Besides binary clinical endpoints, often some time-to-event endpoint, e.g., "time to relapse" will be of interest. The data then is given by $(t_i, \delta_i, x_i), i = 1, \ldots, n$, where t_i is the last observation time, δ_i is an event indicator, taking value 1 if an event occurred at that time and 0 otherwise, and x_i again is a potentially high-dimensional covariate

vector. The Cox proportional hazards model

$$h(t|x_i) = h_0(t)\exp(\eta_i) = h_0(t)\exp(x_i'\beta)$$

then considers the hazard $h(t|x_i)$ at time t, given the covariate values, i.e., the instantaneous risk of having an event, employing an unspecified baseline hazard $h_0(t)$. In a low-dimensional setting, the parameter vector β is estimated by maximizing the partial log-likelihood $l(\beta)$ (for more details see Andersen et al., 1993, for example).

Componentwise likelihood-based boosting is a technique that allows for estimation of both types of models in a high-dimensional setting (see Tutz and Binder, 2007, for generalized linear models, and Binder and Schumacher, 2008b, for the Cox proportional hazards model). Starting with an estimated parameter vector $\hat{\beta}^{(0)} = (0, \ldots, 0)'$, where all covariates have no effect, i.e., none is included in the model, a series of boosting steps $k = 1, \ldots, M$, is performed.

In each boosting step, a candidate model with linear predictor

$$\eta_i^{(k)} = \hat{\eta}_i^{(k-1)} + \gamma_j^{(k)} x_{ij} \quad j = 1, \ldots, p,$$

is considered for each covariate. The information from the previous boosting steps enters via the fixed offset $\hat{\eta}_i^{(k-1)} = x_i'\hat{\beta}^{(k-1)}$ (where for generalized linear models intercept terms $\beta_0^{(k)}$ or $\hat{\beta}_0^{(k-1)}$ have to be included in the candidate models as well as for offset calculation).

The parameters $\gamma_j^{(k)}$ are estimated by maximizing the penalized likelihood

$$l_{pen}(\gamma_j^{(k)}) = l(\gamma_j^{(k)}) + \frac{\lambda}{2}\gamma_j^{(k)},$$

where $\lambda > 0$ is a penalty parameter that determines how far the estimates $\hat{\gamma}_j^{(k)}$ are

shrunken towards zero.

The covariate with index j^* that improves the fit the most then is chosen for the componentwise update

$$\hat{\beta}_j^{(k)} = \begin{cases} \hat{\beta}_j^{(k-1)} + \hat{\gamma}_j^{(k)} & \text{if } j = j^* \\ \hat{\beta}_j^{(k-1)} & \text{otherwise} \end{cases}$$

and $\hat{\eta}_i^{(k)} = x_i' \hat{\beta}^{(k)}$. Specifically, the element j^* of the estimated parameter vector is determined by looking for the covariate that maximizes the penalized log-likelihood $l_{pen}(\gamma_j^{(k)})$ in Tutz and Binder (2007). In Binder and Schumacher (2008b) it is selected via the penalized score statistic

$$U_j^{(k)} (I_j^{(k)} + \lambda)^{-1} U_j^{(k)}, \tag{1}$$

where $U_j^{(k)} = U(0)$ is the score function $U(\gamma) = \partial l(\gamma)/\partial\gamma$, and $I_j^{(k)} = I(0)$ is the Fisher information $I(\gamma) = \partial^2 l(\gamma)/\partial^2\gamma$, both evaluated at zero, however, incorporating the offset. In principle, this could also be employed for determining updates when performing componentwise likelihood-based boosting for generalized linear models and should result in similar estimates compared to selection based on the penalized log-likelihood.

The number of boosting steps M has to be carefully controlled to avoid overfitting. It can, e.g., be chosen by cross-validation. In the resulting models, many of the covariates will not have received any update, i.e., they will not be included. Including only a small number of covariates, e.g., provides a short list of genes deemed informative. The penalty parameter λ is of minor importance, as long as it is chosen large enough, such that the update in each boosting step is small.

2.2 Rank-based p-values

Componentwise boosting is closely related to the lasso. Specifically, in a continuous response setting the estimates from componentwise boosting with small step sizes will be very similar to the estimates from the lasso (Efron et al., 2004). Therefore, componentwise boosting shares also the properties of the lasso concerning groups of correlated covariates with an effect on the response. There, often only one covariate from such a group will receive a non-zero estimate, and the others will not be included in the model (see Zou and Hastie, 2005, for example). This could be expected to result in large variability with respect to the list of covariates with non-zero estimates. Bootstrap inclusion frequencies could provide some information on this. However, our starting point in the present research was to look for more detailed information. Specifically, we considered ways, how such groups of correlated covariates, where only one enters into the final model, could be recognized from the course of the boosting steps. In the end this resulted in the following proposal for obtaining p-values. The resulting development is not only valid for groups of correlated covariates, but will allow for quantifying uncertainty for any covariate.

For obtaining a p-value for each covariate, a statistic and a null distribution is needed, which will be introduced in the following.

2.2.1 Extracting a statistic for each covariate

In each boosting step, the penalized score statistic (1) is calculated for every covariate. The covariate with the largest value of the score statistic is selected for an update, i.e., if it is not yet in the model it will be included. Covariates that are only almost selected in a boosting step can also be recognized from a large value of the score statistic. Therefore, aggregating the value of the score statistic for a covariate over the course of the boosting

steps should provide an indication of importance of that covariate, from the perspective of componentwise boosting.

Unfortunately, the value of the score statistic is difficult to compare between boosting steps, as the values decrease, and more generally depend on many factors such as the overall amount of information in the data. For obtaining values that are comparable between boosting steps, we propose to calculate the rank of the score statistics. Specifically, in each boosting step, the values of the score statistic are ranked. The covariate with the largest rank is the only one that will be updated in this boosting steps. However, other covariates with large ranks can be considered as being close to update/inclusion. The ranks are the same regardless of whether employing the penalized score statistic (1) or the traditional unpenalized score statistic (with $\lambda = 0$).

For judging whether a covariate is often close to inclusion/update, the rank calculated in a boosting step for this covariate has to be aggregated across boosting steps. We suggest to employ the per-covariate median of the rank values across boosting steps. Naturally, other aggregating quantities could be considered. For example, we performed initial experiments using the maximum rank across boosting steps. However, this resulted in a rather problematic distribution. In contrast, as will be seen, the median rank as a statistic for each single covariate allows for reliable inference.

2.2.2 Null distribution

For judging the evidence for effects on the response at the level of the covariates, compared to overall prediction performance, the covariates could, e.g., be ordered according to the median ranks. Compared to ordering based on statistics for univariate tests, this takes the multivariable regression model into account, i.e., the effect of a covariate is judged adjusted for the potential effects of the other covariates. Naturally, this adjustment is only performed to the extent that can be provided by the sparse models with a

linear predictor that are fitted by componentwise boosting.

The median rank of a covariate can be considered as a test statistic for the null hypothesis that this covariate has no effect on the response. Similar to other test statistics, the actual null hypothesis may be larger, as componentwise boosting can detect only some kinds of effects. For example, linearity of the effect and some amount of sparseness are assumed. Furthermore, the number of boosting steps is determined by cross-validation, i.e., the distribution of the median rank values is tied to prediction performance. Therefore, any resulting p-values will be closely connected to the fitted model. However, this is no downside, but is wanted for providing uncertainty information specifically for the fitted model.

For obtaining p-values corresponding to the median rank values as a test statistic, a null distribution is needed. Specifically, the distribution of the median rank needs to be known, given that there is no effect. As componentwise boosting is a rather simple procedure, such a null distribution could certainly be derived analytically for specific settings, e.g., for orthogonal covariates. However, with gene expression data, complex correlation structure between covariates is to be expected, making analytical results difficult to obtain. Therefore, we suggest to employ a permutation-based approach for obtaining a null distribution.

New data sets without covariate effect are generated based on the original data by permuting the response. This leaves the covariate structure intact and corresponds to the general null hypothesis of no connection between the covariates and the response. In each of these data sets, componentwise boosting is performed, using the same number of boosting steps as in the original data (e.g., determined by cross-validation there). Then the median rank for each covariate is determined in each data set.

If cross-validation would be performed in the data sets with permuted response, the selected number of boosting steps mostly would be close to zero, i.e., only few rank

values would be available for each covariate. These would be close to the ranking of a univariate statistic. In contrast, by using the same number of boosting steps as in the original data, values are obtained that reflect the multivariable structure and the assumed level of information of the original model. As an added benefit, computational demand is rather low, e.g., compared to bootstrap inclusion frequencies or the approach of Meinshausen et al. (2009), where all model building steps, including cross-validation, have to be performed in each data set.

The median rank values of all covariates and all data sets with permuted response are then taken together to form an empirical representation of the null distribution. In initial experiments, we alternatively considered forming a separate null distribution for each covariate, for taking the specific correlation structure for this covariate into account. However, this did not provide better results. We also considered different numbers of repetitions with permuted response, and 100 repetitions provided satisfactory results.

The p-value for a covariate then is obtained by taking median rank value for this covariate from the original data and determining the proportion of median rank values from the empirical null distribution that is equal to or smaller than this value.

3 Simulation study

For evaluating the properties of the proposed rank-based p-values for componentwise likelihood-based boosting, we performed a simulation study. Criteria for evaluation are identification of covariates with true non-zero regression parameters and the actual Type I error rate with respect to calling covariates informative, compared to the nominal level given by the p-value for a covariate.

We will consider models for time-to-event data. All approaches will be based on the

Cox proportional hazards model, where the baseline hazard is not estimated. However, omission of covariates from a model might influence performance for the other covariates via the baseline hazard (Gerds and Schumacher, 2001). Therefore, this presents a potentially very difficult setting. If rank-based p-values are seen to perform well there, good performance could also be expected with generalized linear models, e.g., for a binary response.

Rank-based p-values will be obtained by fitting Cox proportional hazards models by componentwise likelihood-based boosting (Binder and Schumacher, 2008b), where the number of boosting steps is chosen by 10-fold cross-validation. For comparison, p-values from univariate Cox models will be considered for each covariate. Unregularized estimation is performed by standard maximum partial likelihood technique, and p-values are obtained from standard Wald tests.

We also consider some multivariable quantities, based on bootstrap data sets. For each data set in the simulation study, 100 bootstrap data sets are generated by drawing $0.632n$ observations without replacement, for avoiding a complexity selection bias (Binder and Schumacher, 2008a). Model-building using likelihood-based boosting is performed in each of these bootstrap data sets, including selection of the number of boosting steps by 10-fold cross-validation. Bootstrap inclusion frequencies (BIFs) are obtained by recording the proportion of bootstrap data sets where a covariate was included in the fitted models. This has been suggested for judging the importance of covariates (see Sauerbrei and Schumacher, 1992; Qiu et al., 2006, for example). Furthermore, the estimated regression parameters for each covariate are averaged over all bootstrap samples, effectively performing frequentist model averaging (Hjort and Claeskens, 2003). This model averaging estimates could be expected to provide more details compared to bootstrap inclusion frequencies.

We consider a fixed number of $n = 200$ observations. Survival times are generated from

a Cox exponential model (see Bender et al., 2005, for example) with baseline hazard 0.1. Censoring times are determined from an exponential distribution with mean 10.

There are $p = 1000$ covariates drawn from a standard normal distribution, where two types of correlation structure are considered. In the first setting, all covariates are uncorrelated. In the second setting, correlation for pairs of covariates with indices j and k is $0.8^{|j-k|}$ for the first 200 covariates, $0.5^{|j-k|}$ for the second block of 200 covariates, and equal to zero for the rest. There are 50 covariates with non-zero effects, which are all located in the first block of 200 covariates at indices $\{1, 4, 7, \ldots, 145, 148\}$, i.e., the maximum correlation between informative covariates is $0.8^3 = 0.512$.

The true regression parameters are determined based on the overall effect size $c_{eff} \in \{0.25, 0.5, 1, 2\}$. The regression parameters for the jth informative covariate then is given by $c_{eff} \cdot c_{decr}^{j}$. Values of $c_{decr} \in \{0.7, 0.9, 0.95, 0.99\}$ provide for a wide range of covariate effect profiles. For example, $c_{decr} = 0.7$ will result in a small number of covariates with large parameter values, while the rest has rather small values. In contrast, $c_{decr} = 0.99$ means that most covariates have a regression parameter of similar size. A total of 20 repetitions is performed for each scenario.

Table 1 shows the performance of the various quantities concerning identification of covariates with non-zero regression parameters. For each quantity, a set of cutoffs was applied for calling a covariate informative. Plotting the proportion of correctly identified covariates at each cutoff against the proportion of covariates erroneously deemed influential results in a ROC curve. The values in Table 1 show the mean area under this curve. The various scenarios are characterized by the standard deviation of the true linear predictors, larger values meaning more information in the covariates with respect to survival times.

For uncorrelated covariates, bootstrap inclusion frequencies consistently result in the worst performance. Better performance compared to univariate p-values is seen only for

Table 1: Mean area under the ROC curve (AUC) for identification of covariates with true non-zero effect, for simulation scenarios with various effect sizes (c_{eff}) and patterns of effect distribution (c_{decr}) for uncorrelated and correlated covariates, characterized by the (mean) standard deviation of the linear predictor (lp). Identification is based on univariate p-values (univ), bootstrap inclusion frequencies (BIF), bootstrap averages of parameter estimates (ave), or rank-based p-values (rank). The largest value in each scenario is highlighted by boldface.

		uncorrelated					correlated				
c_{eff}	c_{decr}	lp	univ	BIF	ave	rank	lp	univ	BIF	ave	rank
0.25	0.7	0.4	0.517	0.516	0.515	**0.520**	0.4	**0.545**	0.515	0.515	0.542
	0.9	0.6	0.558	0.543	0.542	**0.561**	0.6	0.642	0.579	0.581	**0.646**
	0.95	0.8	0.585	0.576	0.575	**0.589**	0.8	0.728	0.661	0.667	**0.741**
	0.99	1.4	0.708	0.640	0.639	**0.709**	1.4	0.891	0.833	0.837	**0.900**
0.5	0.7	0.7	0.541	0.540	0.541	**0.542**	0.7	0.567	0.537	0.536	**0.568**
	0.9	1.2	0.603	0.596	0.597	**0.605**	1.2	0.665	0.622	0.624	**0.686**
	0.95	1.6	0.666	0.655	0.655	**0.677**	1.6	0.761	0.746	0.750	**0.812**
	0.99	2.9	0.770	0.758	0.755	**0.781**	2.9	0.912	0.884	0.885	**0.931**
1	0.7	1.4	0.559	0.556	0.558	**0.565**	1.4	0.572	0.538	0.539	**0.586**
	0.9	2.3	0.631	0.645	0.643	**0.652**	2.3	0.667	0.660	0.661	**0.718**
	0.95	3.3	0.695	0.714	0.709	**0.721**	3.3	0.787	0.793	0.793	**0.850**
	0.99	5.7	0.796	0.789	0.787	**0.814**	5.7	0.913	0.898	0.899	**0.937**
2	0.7	2.8	0.562	0.572	0.571	**0.577**	2.8	0.574	0.555	0.557	**0.603**
	0.9	4.7	0.638	0.666	0.665	**0.676**	4.7	0.669	0.693	0.693	**0.734**
	0.95	6.5	0.709	0.731	0.729	**0.743**	6.5	0.792	0.809	0.809	**0.861**
	0.99	11.4	0.801	0.795	0.795	**0.823**	11.4	0.914	0.899	0.899	**0.937**

a small number of scenarios with correlated covariates and strong effect. However, it would have been expected that fitting of a multivariable model provides an advantage whenever there is correlation between informative covariates. This is seen not to be the case when employing bootstrap inclusion frequencies here. Model averaging estimators also do not seem to perform better.

When employing the proposed rank-based p-values, univariate p-values are outperformed in almost all settings. There is a considerable difference in the scenarios with correlated covariates, i.e., the rank-based p-values leverage the advantage of fitting a multivariable model. The difference is smaller for uncorrelated covariates, where multivariable modeling seems to be less important for identification of informative covariates. Nevertheless, multivariable modeling does not seem to turn into a disadvantage when employing rank-based p-values.

For considering the proposed rank-based p-values useful, not only identification of informative covariates, but also the corresponding Type I error rates have to be investigated. Ideally, when rejecting the null hypothesis of no effect for covariates with p-values smaller than some value α, the proportion of covariates wrongly deemed influential should be approximately equal to α. Conservative behavior, i.e., the proportion being smaller than α, could still be considered acceptable. However, anti-conservative behavior, i.e., the proportion being larger than α, is problematic.

Table 2 shows the mean actual Type I error rates for various nominal levels α, when employing univariate or rank-based multivariable p-values. A nominal level of $\alpha = 0.001$ could be problematic, as there are only 950 covariates without effect in each repetition for calculating the Type I error rate.

In scenarios without correlation, the univariate p-values keep all levels well, which might have been expected. There are some scenarios, where the rank-based p-values are anti-conservative at a level of $\alpha = 0.001$. For all other levels slightly conservative behavior

Table 2: Mean of the actual Type I error rates (standard deviation in parentheses) compared to the nominal level for univariate p-values and rank-based p-values from componentwise boosting, for scenarios with various effect sizes (c_{eff}) and patterns of effect distribution (c_{decr}) for uncorrelated and correlated covariates.

		nominal level					
		univariate p-values			rank-based p-values		
c_{eff}	c_{decr}	0.001	0.01	0.1	0.001	0.01	0.1
		uncorrelated covariates					
0.25	0.7	0.001 (0.001)	0.011 (0.004)	0.101 (0.010)	0.001 (0.001)	0.009 (0.001)	0.098 (0.002)
	0.9	0.001 (0.001)	0.009 (0.003)	0.101 (0.007)	0.001 (0.001)	0.008 (0.001)	0.096 (0.002)
	0.95	0.001 (0.001)	0.010 (0.003)	0.099 (0.010)	0.001 (0.001)	0.008 (0.001)	0.094 (0.003)
	0.99	0.001 (0.001)	0.010 (0.003)	0.098 (0.009)	0.001 (0.001)	0.006 (0.001)	0.087 (0.003)
0.5	0.7	0.001 (0.001)	0.011 (0.003)	0.102 (0.010)	0.001 (0.001)	0.007 (0.001)	0.096 (0.002)
	0.9	0.001 (0.001)	0.010 (0.003)	0.100 (0.013)	0.001 (0.001)	0.006 (0.002)	0.092 (0.002)
	0.95	0.001 (0.001)	0.010 (0.003)	0.100 (0.008)	0.001 (0.001)	0.005 (0.002)	0.087 (0.002)
	0.99	0.001 (0.001)	0.011 (0.003)	0.101 (0.007)	0.001 (0.001)	0.005 (0.001)	0.081 (0.002)
1	0.7	0.001 (0.001)	0.011 (0.004)	0.101 (0.012)	0.003 (0.001)	0.007 (0.001)	0.094 (0.003)
	0.9	0.001 (0.001)	0.011 (0.003)	0.102 (0.009)	0.003 (0.001)	0.006 (0.001)	0.088 (0.002)
	0.95	0.001 (0.001)	0.010 (0.003)	0.101 (0.006)	0.001 (0.001)	0.005 (0.001)	0.083 (0.003)
	0.99	0.001 (0.001)	0.010 (0.003)	0.098 (0.005)	0.001 (0.001)	0.005 (0.001)	0.078 (0.003)
2	0.7	0.001 (0.001)	0.010 (0.004)	0.103 (0.013)	0.005 (0.001)	0.009 (0.001)	0.091 (0.003)
	0.9	0.001 (0.001)	0.010 (0.003)	0.103 (0.009)	0.002 (0.001)	0.005 (0.001)	0.085 (0.003)
	0.95	0.001 (0.001)	0.010 (0.003)	0.101 (0.010)	0.002 (0.001)	0.005 (0.001)	0.081 (0.002)
	0.99	0.001 (0.001)	0.011 (0.003)	0.100 (0.007)	0.001 (0.001)	0.005 (0.002)	0.078 (0.003)
		correlated covariates					
0.25	0.7	0.005 (0.002)	0.016 (0.004)	0.109 (0.011)	0.001 (0.001)	0.007 (0.001)	0.096 (0.003)
	0.9	0.012 (0.004)	0.028 (0.004)	0.124 (0.010)	0.001 (0.001)	0.005 (0.002)	0.090 (0.003)
	0.95	0.018 (0.005)	0.040 (0.006)	0.144 (0.012)	0.001 (0.001)	0.005 (0.002)	0.082 (0.003)
	0.99	0.025 (0.008)	0.058 (0.009)	0.174 (0.012)	0.000 (0.001)	0.004 (0.002)	0.074 (0.003)
0.5	0.7	0.009 (0.003)	0.021 (0.003)	0.115 (0.011)	0.001 (0.001)	0.006 (0.001)	0.094 (0.002)
	0.9	0.020 (0.004)	0.037 (0.006)	0.131 (0.011)	0.001 (0.001)	0.005 (0.002)	0.084 (0.003)
	0.95	0.027 (0.005)	0.048 (0.006)	0.149 (0.012)	0.001 (0.001)	0.004 (0.001)	0.077 (0.002)
	0.99	0.035 (0.008)	0.071 (0.009)	0.180 (0.014)	0.001 (0.001)	0.004 (0.002)	0.069 (0.003)
1	0.7	0.011 (0.003)	0.022 (0.004)	0.119 (0.012)	0.002 (0.001)	0.005 (0.001)	0.091 (0.003)
	0.9	0.023 (0.004)	0.038 (0.006)	0.132 (0.010)	0.002 (0.001)	0.005 (0.002)	0.082 (0.003)
	0.95	0.032 (0.005)	0.053 (0.007)	0.151 (0.008)	0.001 (0.001)	0.004 (0.002)	0.074 (0.003)
	0.99	0.039 (0.008)	0.071 (0.008)	0.182 (0.014)	0.000 (0.001)	0.004 (0.002)	0.068 (0.002)
2	0.7	0.012 (0.002)	0.023 (0.005)	0.116 (0.011)	0.003 (0.001)	0.007 (0.001)	0.089 (0.002)
	0.9	0.024 (0.003)	0.038 (0.006)	0.129 (0.013)	0.002 (0.001)	0.005 (0.002)	0.080 (0.003)
	0.95	0.033 (0.005)	0.054 (0.007)	0.154 (0.011)	0.001 (0.001)	0.004 (0.002)	0.073 (0.004)
	0.99	0.039 (0.008)	0.072 (0.009)	0.183 (0.012)	0.000 (0.001)	0.004 (0.002)	0.069 (0.002)

is seen. There is a small difference between univariate p-values and rank-based p-values with respect to variability of the Type I error rates. The rank-based approach seems to result in somewhat less variability.

For settings with correlated covariates, the univariate p-values are seen to result in anti-conservative behavior, with large deviations from the nominal level when there is more information in the covariates. This might have been expected, as the univariate approach can not take the overall correlation structure into account. In contrast, the rank-based p-values are based on a multivariable approach. A seen from Table 2, this seems to pay off. The actual Type I error rates are similar to the uncorrelated scenarios, i.e., the approach is not affected much by correlation structure. There is still some anti-conservative behavior for $\alpha = 0.001$. The tendency towards conservative behavior seems to be increased for $\alpha = 0.01$ and $\alpha = 0.1$. Variability of the Type I error is increased for univariate p-values, corresponding to the increased overall level. For the rank-based p-values, not systematic increase or decrease can be seen, compared to scenarios without correlation.

4 Application example

Rosenwald et al. (2002) analyzed the data of 240 patients with diffuse large-B-cell lymphoma, for linking gene expression to time to death. There were 138 deaths observed, with a five year overall survival of 48%. Gene expression at baseline is given via $p = 7399$ microarray measurements. An established clinical predictor, the International Prognostic Index (IPI — a combination of five clinical features), is available for $n = 222$ patients, which will be considered for analysis in the following. Extensive reanalysis of this data has already been performed, employing sparse estimates of Cox proportional hazards models (see Segal, 2006, for example). Componentwise likelihood-based boosting

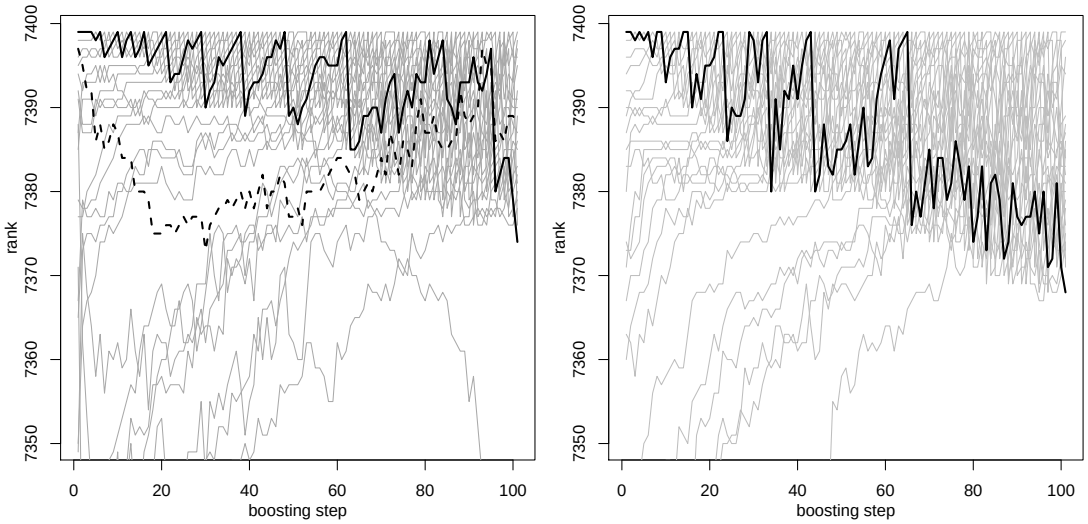

Figure 1: Rank of microarray covariates selected for the final model in the course of the boosting steps (grey curves) for the original data (left panel) and one data set with permute response (right panel). The covariates selected in the first boosting step in the respective data sets are highlighted by solid black curves. Also, the values for microarray covariate X15804 (not selected in the original data) are given (dashed black cure).

of the Cox model for this data has been considered in Binder and Schumacher (2008b), where good prediction performance was demonstrated (more details can be found there). Specifically, the boosting approach can easily incorporate the IPI in a mandatory, unregularized way, allowing for reasonable comparison to a purely clinical model, containing only the IPI. Such a combined model, including the clinical covariate IPI as well as the microarray covariates, will be considered in the following, but only quantification of uncertainty for the microarray covariates will be considered.

The number of boosting steps is selected by 10-fold cross-validation, resulting in $M = 101$ steps with 21 non-zero parameter estimates, i.e., genes deemed informative. Rank-based p-values were calculated as described in Section 2.2, using 100 data sets with permuted response. Using the same number of boosting steps $M = 101$ as in the original data, a median of 26 covariates was included.

Figure 1 shows the rank of the 21 covariates selected in the original data in the course of the boosting steps (left panel) and the rank of the 25 selected covariates in one of the data sets with permuted response (right panel). For illustration, the respective covariates selected in the first boosting step in the two data sets are indicated by a solid black curves. As can be seen, the rank of these covariates drops only slightly after having been selected in a boosting step (corresponding to a rank of 7399). Several times, the rank increases again, until the covariates are selected one more time. This pattern seems to be frequent for the selected covariates, as indicated by the triangular area with high curve density. Consequently, the overall level of the rank curves, represented by the median, appropriately reflects the importance of a covariate, as there is, e.g., no drastic drop off. The dashed black line shows the values for the microarray covariate with GenBank accession number X15804, which has been deemed informative by Rosenwald et al. (2002), but is not included in the model fitted by componentwise boosting. This covariate receives a large rank in early boosting steps, i.e., it is close to selection. The values then decrease, potentially because covariates have been included that contain similar information. However, the decrease is not strong, and eventually the rank increases again, coming close to selection. Correspondingly, the median of the values will reflect that this covariate is close to selection over the whole course of boosting steps, i.e., that it might be important. Similar patterns for rank curves are seen for data sets with permuted response, i.e., under the null hypothesis (right panel of Figure 1). For example, there also is no sudden drop off for selected covariates. Therefore, the median also seems to be an appropriate summary under the null hypothesis, providing the basis for calculating p-values.

For comparison, p-values were calculated from separate univariate Cox proportional models for each microarray covariate, where the IPI was included in each model in addition. With respect to multivariable models, two quantities are considered for comparison. Since the microarray covariates have been centered and standardized before applying

Table 3: Criteria for judging microarray features (GenBank accession no. marked by "*" if also reported by Rosenwald et al., 2002), sorted by the values of the rank-based p-values for the multivariable model fitted by boosting: rank of the univariate p-values (univ.rank), estimated regression parameters for componentwise boosting (param), and bootstrap inclusion frequencies (BIF).

Rank	GenBank	univ.rank	param	BIF	Rank	GenBank	univ.rank	param	BIF
1	BC012161	2	0.194	0.77	22	M20430*	32	-	0.03
2	X77743	1	-0.198	0.89	23	-	21	-	0.09
3	U15552	3	0.119	0.57	24	NM_000176	23	-	0.03
4	AB018289	11	-0.045	0.30	25	AK021632	504	-0.018	0.10
5	AF127481	37	0.112	0.48	26	X00452*	19	-	0.06
6	D42043	12	-0.086	0.49	27	L47345	38	-	0.16
7	AF075587	10	-0.031	0.15	28	U12979	9	-	0.15
8	M26004	65	-0.044	0.13	29	M16276	95	-	0.03
9	K01171*	7	-0.098	0.16	30	X87241	170	-0.008	0.11
10	X70649	8	0.019	0.24	31	M26004	164	-	0.02
11	-	18	-0.020	0.15	32	X00452*	29	-	0.03
12	BF129543	4	-0.043	0.34	33	M20430*	28	-	0.03
13	U46767	15	0.019	0.18	34	L13616	25	-	0.03
14	-	20	0.028	0.23	35	M32110	5	-	0.09
15	NM_000176	16	-	0.16	36	-	83	-	0.11
16	X15804*	6	-	0.19	37	U47414	63	-	0.03
17	-	54	0.019	0.11	38	M26004	135	-	0.02
18	-	79	-	0.13	39	M26004	152	-	0.04
19	AF189009	48	-0.019	0.18	40	AF189009	45	-	0.09
20	AF041261	36	0.018	0.17	41	X00452*	31	-	0.02
21	M16276	87	-0.009	0.14	42	M23452	92	0.009	0.10

componentwise boosting, the absolute value of estimated regression parameters might indicate importance. Furthermore, we determined bootstrap inclusion frequencies. Following Binder and Schumacher (2008a), 100 bootstrap data sets were generated by drawing $0.632n$ observations without replacement for each data set, for avoiding a complexity selection bias. All model building steps for componentwise boosting, including selection of the number of boosting steps, were performed in each bootstrap sample, resulting in 100 estimated parameter vectors. The bootstrap inclusion frequency is calculated as the proportion of bootstrap data sets where a covariate receives a non-zero parameter estimate.

The criteria for judging importance of genes with respect to survival are given in Table

3 for a selected number of genes. These were selected such that all genes are given that have a rank-based p-value smaller or equal to the gene with the largest rank-based p-values that received a non-zero parameter estimate from boosting in the original data. The genes are ranked by the value of the rank-based p-values.

Where available, the GenBank accession number is given. It is marked by a "*" if it is in the list of 17 representative predictive genes given in Table 2 of Rosenwald et al. (2002). In the latter analysis, univariate p-values were calculated without adjusting for the IPI. Therefore, some differences to the present analysis are to be expected. Seven of the genes identified by Rosenwald et al. (2002) are recovered among the top ranked 42 genes according to the rank-based p-values. For comparison, eight genes could be recovered if choosing the 42 best genes according to the univariate p-values, adjusted for the IPI.

For the univariate p-values, the ranks are given in Table 3, allowing for judging agreement of ranking according to univariate and the proposed multivariable model p-values. Both approaches are seen to agree with respect to the three top-ranked genes. However, there is a considerable number of genes that is deemed more informative by the multivariable p-values and vice versa.

There is a closer agreement between the genes deemed important by the proposed rank-based p-values and the bootstrap inclusion frequencies. Nevertheless, several of the genes deemed important by Rosenwald et al. (2002) receive relatively low bootstrap inclusion frequencies and are only highlighted by the rank-based p-values.

Overall, all three quantities, univariate p-values, bootstrap inclusion frequencies, and rank-based p-values, point out several genes that are reported by Rosenwald et al. (2002), but were not included in the model fitted by componentwise boosting to the original data. However, there is a tendency of smaller rank-based p-values corresponding to larger values of the estimated regression parameters. Therefore, the former might be

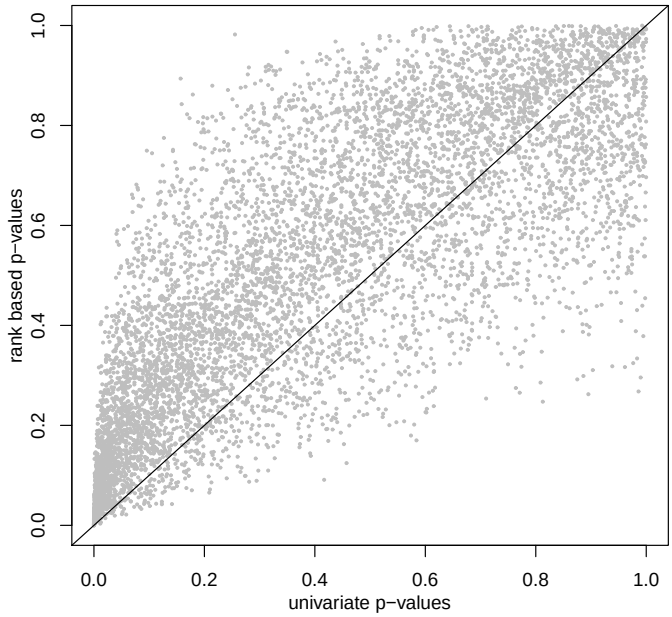

Figure 2: Univariate p-values (adjusted for the effect of the IPI) and rank-based p-values from a multivariable model fitted by componentwise boosting.

considered a good supplement to the latter.

Figure 2 provides a more detailed comparison of the univariate p-values and the rank-based p-values for the multivariable model fitted by componentwise boosting. Both are given for each of the $p = 7399$ microarray covariates. The rank-based p-values are systematically larger compared to the univariate p-values. In particular, the difference is large for small p-values, which are critical for picking out a small number of informative genes. Such a difference between the two types of p-values has already been seen in the simulation study, where the univariate p-values showed anti-conservative behavior. Therefore, the rank-based p-values could be considered more reliable in this application example.

5 Discussion

There are several techniques for the fitting of risk prediction models to high-dimensional molecular data. Approaches such as the lasso or componentwise likelihood-based boosting allow for fitting of linear models, where the contribution of each covariate is represented by one regression parameter. While parameter estimates equal to zero or not equal to zero indicate which covariates might be important, this does not provide information on uncertainty. Bootstrap inclusion frequencies provide some information on the latter, but do not allow for the standard probability interpretation that is readily available from univariate test-based approaches.

To remedy this, we proposed an approach for obtaining p-values for models fitted by componentwise likelihood-based boosting. This does not rely only on the fitted models, but extracts information from the fitting process. Ranking the values of the score statistic in each boosting step for every covariate indicates not only which covariates are included into the model or updated, but highlights candidates near to inclusion. Aggregating this over the course of the boosting step provides information on covariate importance in the context of componentwise likelihood-based boosting. For obtaining p-values, a null distribution was obtained by repeating the fitting process on data with permuted response.

We performed a simulation study in the challenging setting of high-dimensional time-to-event data. While bootstrap inclusion frequencies were seen to provide worse covariate identification performance compared to univariate techniques, even with correlated covariates, the rank-based p-values performed best in most of the scenarios. As might have been expected, the largest gain was seen for scenarios with correlated covariates.

Inspecting the nominal and the actual Type I error rates, univariate p-values were seen to exhibit anti-conservative behavior with correlated covariates. In contrast, the rank-based

p-values keep their error level in most settings, showing slightly conservative behavior. In addition, variability of Type I error rates was smaller when employing rank-based p-values.

A similar pattern was seen in the application example with high-dimensional gene expression data and a time-to-event endpoint. The univariate p-values were systematically larger compared to the rank-based p-values. Based on the simulation results, the latter seem to be more reliable. When comparing genes deemed important by various criteria, it was seen that there is considerable agreement between univariate p-values and the rank-based p-values, and also between the latter and bootstrap inclusion frequencies. At least, the rank-based p-values were seen to provide a reasonable ranking for the genes selected by componentwise boosting. In addition, some genes that potentially might have been missed were indicated.

In summary, the idea of leveraging information from a stagewise model fitting process has been seen to be successful in providing p-values for models fitted by componentwise likelihood-based boosting. This idea might be more generally applicable. Many approaches for fitting models to high-dimensional data, such as random forests (Breiman, 2001), employ a large number of small steps. By systematically leveraging information from this process, reasonable quantification of uncertainty could potentially be obtained. However, this is the subject of future research.

While currently being limited to componentwise boosting, the proposed approach at least is a first step towards better quantification of uncertainty in high-dimensional risk prediction models, providing a model fitting technique where p-values for single model components are available in addition to good prediction performance. This promises to considerably improve statistical modeling practice when linking high-dimensional molecular data to clinical endpoints.

Acknowledgements

We gratefully acknowledge support from Deutsche Forschungsgemeinschaft (DFG Forschergruppe FOR 534).

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, Berlin.

Baek, S., Tsai, C., and Chen, J. (2009). Development of biomarker classifiers from high-dimensional data. *Briefings in Bioinformatics*. Doi: 10.1093/bib/bbp016.

Baker, S. G. and Kramer, B. S. (2006). Identifying genes that contribute most to good classification in microarrays. *BMC Bioinformatics*, 7:407.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300.

Binder, H. and Schumacher, M. (2008a). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Statistical Applications in Genetics and Molecular Biology*, 7(1):Article 12.

Binder, H. and Schumacher, M. (2008b). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9:14.

Boulesteix, A.-L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10(5):556–68.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

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.

Gerds, T. A. and Schumacher, M. (2001). On functional misspecification of covariates in the Cox regression model. *Biometrika*, 88:572–580.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Kaderali, L., Zander, T., Faigle, U., Wolf, J., Schultze, J. L., and Schrader, R. (2006). CASPAR: A hierarchical bayesian approach to predict survival times in cancer from gene expression data. *Bioinformatics*, 22(12):1495–502.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics*, 34(5):2555–2591.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, U.K., 2nd edition.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. arXiv:0811.2177v2 [stat.ME]. To appear in Journal of the American Statistical Association.

Noma, H., Matsui, S., Omori, T., and Sato, T. (2009). Bayesian ranking and selection methods using hierarchical mixture models in microarray studies. *Biostatistics*. Doi: 10.1093/biostatistics/kxp047.

Park, M. Y. and Hastie, T. (2007). L_1-regularization path algorithms for generalized linear models. *Journal of the Royal Statistical Society B*, 69(4):659–677.

Qiu, X., Xiao, Y., Gordon, A., and Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, 7:50.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyna, R. D., Muller-Hermelink, H. K., Smeland, E. B., and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, 346(25):1937–1946.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, 11(16):2093–2190.

Schumacher, M., Binder, H., and Gerds, T. A. (2007). Assessment of survival prediction models based on microarray data. *Bioinformatics*, 23(14):1768–1774.

Segal, M. (2006). Microarray gene expression data with linked survival phenotypes: Diffuse large-B-cell lymphoma revisited. *Biostatistics*, 7(2):268–285.

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. (2007). Boosting ridge regression. *Computational Statistics & Data Analysis*, 51(12):6044–6059.

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.