Resampling Approaches in Biometrical Applications: Developments in Random Forests and in Bootstrap-based Procedures

Silke Janitza

München 2016

Resampling Approaches in Biometrical Applications: Developments in Random Forests and in Bootstrap-based Procedures

Silke Janitza

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Silke Janitza
aus Troisdorf

München, den 12. Januar 2016

Erstgutachter: Prof. Dr. Anne-Laure Boulesteix

Zweitgutachter: Prof. Dr. Matthias Schmid

Tag der mündlichen Prüfung: 29.04.2016

Summary

In this thesis new insights and developments on resampling approaches are provided. While the first two parts provide new developments on the resampling method random forests, the third and fourth parts investigate bootstrap-based approaches in which either hypothesis testing or model selection are performed on each bootstrap sample.

Random forests are an ensemble of classification or regression trees with each tree being built from a sample drawn either with or without replacement from the original data. While classification and regression problems using random forest methodology have been extensively investigated in the past, there seems to be a lack of literature on handling ordinal regression problems, that is, if response categories have an inherent ordering. In the first part, this thesis investigates if incorporating the ordering information in random forests improves prediction and variable selection. The second part focuses on the identification of relevant variables in high-dimensional settings through the use of random forest's variable importance measures. When using random forest's variable importance measures, the researcher faces the problem that there is no natural cutoff for importance scores that can be used to differentiate between important and non-important variables. Several approaches, such as approaches based on hypothesis testing through permutation-based procedures, have been developed for addressing this problem. While for low-dimensional settings the existing permutation-based approaches might be computationally tractable, for high-dimensional settings typically including several thousands of variables, computing time is enormous. This thesis introduces a computationally fast heuristic variable importance test for high-dimensional data settings.

Other resampling approaches, which are based on the bootstrap, are investigated in the third and fourth parts of this thesis. These address for example stability investigations. Repeating the same analysis on a large number of data samples from the same data generating process allows one to draw conclusions on how stable the results are against data perturbations. Since in practical applications the data generating process is unknown, several authors proposed using the bootstrap instead. However, applying the data analysis on bootstrap samples as if they were samples drawn from the true distribution might be misleading if the data analysis includes hypothesis testing or model selection steps using information criteria or data splitting approaches. This is addressed in the third and fourth parts of this thesis, respectively, and promising solutions are investigated.

Zusammenfassung

In dieser Arbeit werden im ersten und zweiten Teil neue Entwicklungen der Resampling Methode Random Forests vorgestellt, während der dritte und vierte Teil Bootstrap-basierte Verfahren behandelt, in denen Hypothesentests oder Modellselektionsverfahren auf Bootstrap-Stichproben angewendet werden.

Bei der Random Forests Methode handelt es sich um ein Ensemble von Klassifikations- oder Regressionsbäumen, die auf Bootstrap-Stichproben angepasst werden. Während Random Forests sich als beliebtes Verfahren für Klassifikations- und Regressionsprobleme etabliert hat, ist dessen Verwendung in Datensituationen, in denen die Zielgröße ordinal ist, bislang noch nicht hinreichend untersucht. Inwiefern die Information über die Reihenfolge der Response-Kategorien die Güte der Prädiktion und Variablenselektion zu verbessern vermag, wird im ersten Teil dieser Arbeit untersucht. Eine besondere Herausforderung stellt die Identifizierung relevanter Variablen über die in Random Forests integrierten Variablenwichtigkeitsmaße dar. Anhand dieser Maße kann für jede Variable ein Score berechnet werden, der die Wichtigkeit der Variable reflektiert und eine Anordnung der Variablen nach Relevanz ermöglicht. Jedoch existiert kein Richtwert, der die Trennung von relevanten und nicht-relevanten Variablen anhand ihrer Scores erlaubt. Verschiedene Lösungsansätze wie Hypothesentest-basierte Ansätze, die Gebrauch von computational aufwändigen Permutations-Verfahren machen, wurden in der Literatur vorgeschlagen. Während die existierenden Permutations-basierten Testansätze für niedrigdimensionale Daten computational zu bewältigen sind, ist die Rechenzeit bei hochdimensionalen Daten, die nicht selten mehrere tausend Variablen umfassen, enorm. Im zweiten Teil dieser Arbeit wird ein heuristisches, computational effizientes Testverfahren für die Variablenwichtigkeit vorgeschlagen, das für hochdimensionale Daten geeignet ist.

Im dritten und vierten Teil der Arbeit werden spezielle Bootstrap-basierte Verfahren untersucht bei denen Hypothesentests, Informationskriterien oder Kreuzvalidierung auf Bootstrap-Stichproben zur Anwendung kommen. Über die wiederholte Durchführung von statistischen Analysen auf Bootstrap-Stichproben können beispielsweise Kenntnisse über die Stabilität der Ergebnisse gewonnen werden. Jedoch ist einigen Studien nach die Anwendung von Hypothesentests oder Modellselektion über Informationskriterien oder Kreuzvalidierung auf Bootstrap-Stichproben problematisch. Ob und inwiefern die

Ergebnisse spezieller Bootstrap-basierter Verfahren beeinflusst sind, sowie die Ursachen und vielversprechende Lösungsansätze werden im dritten und vierten Teil dieser Arbeit untersucht.

Danksagung

Die vorliegende Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie der Ludwig-Maximilians-Universität München.
Mein tiefster Dank gebührt meiner Doktormutter Prof. Dr. Anne-Laure Boulesteix, die mir die wissenschaftliche Arbeit auf dem interdisziplinären Gebiet der Biostatistik ermöglicht und mir damit meinen langjährigen Wunsch erfüllt hat, auf den ich stets hingearbeitet habe. Für ihr entgegengebrachtes Vertrauen und dem damit verbundenen wissenschaftlichen Freiraum, sowie für die Förderung meiner Ideen bin ich zutiefst dankbar.
Mein Dank gilt außerdem Herrn Prof. Dr. Matthias Schmid, der sich dazu bereit erklärt hat, die Aufgabe des zweiten Berichterstatters zu übernehmen.
Für einen schönen Ausgleich zur wissenschaftlichen Arbeit und seine stets lockere und gelassene Art danke ich Heiko Winter. Schließlich möchte ich meinen Eltern danken, die mir das Statistik-Studium ermöglicht haben, auf dem diese Promotion basiert.

München, im Januar 2016 *Silke Janitza*

Contents

1. Introduction

Bootstrap and bootstrap aggregating

The bootstrap method proposed by Efron and Tibshirani (1993) has become a popular tool that is applied in diverse areas and for different purposes. In the case of the non-parametric bootstrap one draws with replacement from the original data to obtain a bootstrap sample. Repeating this sampling procedure B times results in B bootstrap samples that arise from the original data. On each bootstrap sample the statistic of interest can be computed. Having realizations of the statistic one can for example compute the variability of the statistic, a quantile of interest, a confidence interval or simply approximate the whole underlying distribution of the statistic.

The bootstrap has also been used in completely different contexts. A popular application field of the bootstrap is the combination of multiple classifiers with each classifier being built on a bootstrap sample. This procedure is termed "bagging", short for *bootstrap aggregating*, where the word "aggregating" refers to the fact that the predictions by the B classifiers are aggregated in some way to obtain a more precise prediction (Breiman; 1996a). In the most simple case when the response is numeric, the predictions are averaged. Bagging has been shown to achieve better prediction accuracies than single classifiers in some cases, particularly if the method for constructing classifiers is unstable (see, e.g., Dietterich; 2000). An example for unstable classifiers are decision trees. Decision trees were often used for bagging. Such procedures have lead to the development of the famous random forests method (Breiman; 2001) which is nowadays widely applied.

Random forests

In contrast to bagging trees, in random forests only a random subset of the predictor variables is considered for each split in a tree. This makes the trees more diverse and leads to better predictions. Random forests can be applied for classification (in the case of a nominal response) as well as for regression tasks (in the case of a numeric response). In contrast to many classical statistical methods, they can even be applied in the statistically challenging high-dimensional data setting in which the number of variables, p, is larger

than the number of observations, n. This makes random forests especially attractive for complex high-dimensional molecular data applications. A further advantage of random forests is that they offer so-called variable importance measures that can be used to rank variables according to their predictive abilities.

For nominal and numeric response the application of random forests has been well investigated. However, in some applications the response is neither nominal nor numeric, but something in between; such variables are termed ordinal as their categories have an inherent ordering but the distances between categories cannot be quantified. Examples of ordinal responses in biometrical applications are tumor stages I - IV, disease severity, for example from mild to moderate to severe disease state, and artificially created scores combining several single measurements into one summary measure, like the Apgar score, which is used to assess the health of a newborn. In the case of ordinal response there is no standard random forests procedure. While in the classical random forest algorithm of Breiman (2001) the ordering of a predictor is taken into account by allowing splits only between adjacent categories, the ordering information in the response is ignored (i.e., the response is treated as a nominal variable), and an ensemble of classification trees is constructed. However, ignoring the ordering information results in a loss of information. The question which needs to be addressed is whether predictions by random forests improve by using this information, or not. In the context of variable ranking by random forest's variable importance measures, the question arises of whether variable rankings are more accurate if taking the ordering of the response levels into account when computing the variables' importance scores. Both questions have not been addressed in the literature so far.

The second issue which is investigated in this thesis is on the use of variable importance measures for identifying relevant variables from high-dimensional data. In high-dimensional genomic data often the identification of relevant genes is of interest to gain valuable insights into the functionality and mechanisms that lead to a specific disorder. Moreover, the identification of relevant genes aids in the diagnosis of certain disorders. The random forest method and their implemented variable importance measures have often been used for the identification of biomarkers (e.g., Reif et al.; 2009; Wang-Sattler et al.; 2012; Yatsunenko et al.; 2012). A drawback of the variable importance measures is that there is no natural cutoff for the importance score that can be used to select variables which are likely relevant. Every researcher working with random forest's variable importance measures thus faces the problem where to set this cutoff. Several approaches, for example approaches based on hypothesis testing, have been developed for addressing this problem. The existing testing approaches are permutation-based and require the repeated computation of forests. While for low-dimensional settings those permutation-

based approaches might be computationally tractable, for high-dimensional settings typically including thousands of genes, computing time is enormous and a fast implementation of a variable importance test might be desirable.

Problems related to the bootstrap

Despite its wide applicability, there are situations in which the application of the bootstrap is problematic; Andrews (2000) and Abadie and Imbens (2008) for example show the failure of the bootstrap in two specific situations, and Chernick (2008) (Chapter 9) and Bickel and Freedman (1981) give a more broad overview of a range of problems encountered with the bootstrap. This thesis focuses on two specific problems related to the bootstrap, which deserve further attention: the application of hypothesis tests on bootstrap samples and the application of model selection strategies through the use of information criteria or cross-validation performed on bootstrap samples.

Hypothesis testing on bootstrap samples

Recently some approaches have been proposed in the biometrical field where hypothesis testing is performed on a bootstrap sample as if it were the original sample. In the statistics and bioinformatics literature the p-values computed from bootstrap samples have been used for example for ranking genes with respect to their differential expression (Mukherjee et al.; 2003), for estimating the variability of p-values (Boos and Stefanski; 2011) and for model stability investigations (Chen and George; 1985; Altman and Andersen; 1989; Sauerbrei and Schumacher; 1992). For the likelihood ratio test (Bollen and Stine; 1992) and for the χ^2-test (Strobl et al.; 2007) it was shown that p-values computed on bootstrap samples do not represent what would be obtained on the original data or new data drawn from the overall population. Other tests might be similarly affected. The consequences for random forests, for example, was a biased split selection (Strobl et al.; 2007). However, the practical impact on many other bootstrap-based procedures relevant to biometrical applications has not yet been studied.

Strobl et al. (2007) recommended using subsampling instead of bootstrapping in random forests to avoid biased split selection. The subsampling procedure, also known as delete-d jackknife (Wu; 1986), is closely related to the bootstrap, but in contrast to the bootstrap, a subsample is created by drawing m observations, with $m < n$, without replacement from the original sample. The subsampling technique has been investigated in the literature and also compared to the bootstrap (Shao and Wu; 1989; Politis and Romano; 1994; Politis et al.; 1999; Hartigan; 1969). It shows asymptotic consistency in cases where the bootstrap fails (Davison et al.; 2003; Chernick; 2008). In particular the type I

error is not increased for test statistics computed on subsamples. For many of the existing bootstrap-based procedures it has not been investigated so far whether subsampling is a useful alternative to the bootstrap.

Error estimation of a prediction modeling strategy by the bootstrap

Bootstrapping is commonly used for the estimation of the error of a prediction modeling strategy as an alternative to, say, cross-validation. For a large number of bootstrap samples drawn from the original data set, a prediction model is fit to the bootstrap sample using the considered strategy and is then used to make predictions for the observations which were not included in this bootstrap sample and are thus considered test data. This yields an estimate for the prediction error of the model and the estimates from all bootstrap samples are averaged.

Many statistical methods involve tuning parameters that must be chosen. An example are gradient boosting methods. These methods combine weak learners in an iterative fashion to obtain a strong learner with high prediction accuracy. The prediction accuracy depends highly on the number of iterations, also called the number of boosting steps. With too many boosting steps, many weak learners are created and the resulting strong learner might be overfit to the data and thus have poor prediction accuracy on new data. If the number of boosting steps is too small, the number of weak learners might be too small to appropriately model the relationship between the covariates and the response. Thus the number of boosting steps has to be carefully chosen, for example through internal cross-validation performed on the bootstrap sample. Binder and Schumacher (2008), however, showed that the resulting error estimate is biased, and that subsampling yields less biased error estimates. The reasons for this bias are unknown and remain to be investigated to aid the development of alternative strategies for avoiding this bias. Alternatively, instead of cross-validation, information criteria may be used for selecting optimal values for tuning parameters. In different contexts Wagenmakers et al. (2004) and Steck and Jaakkola (2003) showed that information criteria derived from bootstrap samples systematically deviate from information criteria derived from original samples. The practical consequences of this systematic deviation on prediction modeling strategies have not yet been explored, and promising alternatives, such as subsampling, remain to be investigated.

Guideline through the thesis

The main part of this thesis consists of five chapters which are related to each other, but are kept self-contained. A background on the random forest methodology is given in

Chapter 2. This chapter mostly surveys current random forest methodology and might be consulted for technical details on tree construction. Chapters 3 and 4 deal with improvements and new methods for random forests. While Chapter 3 shows extensive studies on the appropriate handling of ordinal responses using existing and new approaches, Chapter 4 investigates the performance of a new computationally fast test for random forest's variable importance. Chapters 5 and 6 investigate problems related to the bootstrap from a theoretical and practical point of view. These problems relate to hypothesis testing or model selection through information criteria or data splitting approaches performed on bootstrap samples. A special emphasis is laid on the consequences in biometrical applications, and the use of subsampling is investigated as an alternative. Summaries of the chapters are given in the following.

Chapter 2: Trees, bagging and random forests

In this chapter, the concepts of recursive partitioning and of two popular ensemble methods, bagging and random forests, are briefly described. Impurity-based and test-based split criteria for partitioning the feature space, are reviewed. The bootstrap method on which the ensemble method bagging builds upon, is briefly described. The basic principles of the random forest method which is deeply rooted in the bagging algorithm, are outlined. A special emphasis is put on random forest's out-of-bag observations and its variable importance measures. An application example of random forests in medicine is given at the end.

Chapter 3: Random forests for ordinal responses

This chapter investigates the use of ordinal regression trees developed by Hothorn, Hornik and Zeileis (2006) as base classifiers in the random forest algorithm. In contrast to classification trees, ordinal regression trees make use of the ordering of the response levels. Moreover, two novel permutation variable importance measures are presented which take the ordering of the response levels into account. Extensive simulations and real data-based studies are conducted to investigate whether taking the ordering of the response levels into account leads to more accurate predictions and predictor rankings by random forests.

Chapter 4: A variable importance test for high-dimensional data

In this chapter, a novel heuristic variable importance test is presented that is particularly suitable for high-dimensional molecular data where typically a small subset of the variables carries most or all of the information. In contrast to existing approaches, the test is not permutation-based and thus computationally very fast.

Studies with high-dimensional data are performed to investigate the properties of the test. Moreover, the performance of the new test is compared to the performance of a popular permutation-based testing approach.

Chapter 5: Hypothesis testing on bootstrap samples

In this chapter, it is theoretically and empirically shown that the type I error is increased when performing a Z-test on bootstrap samples. Empirical evidence for an increased type I error is also given for the likelihood ratio test. Further, the distributions of bootstrapped p-values are studied for both the Z-test and the likelihood ratio test. For three bootstrap-based approaches the consequences of the increased type I error are illustrated using a real data application. The use of subsampling is investigated as possible solution.

Chapter 6: Model selection through information criteria and data splitting approaches on bootstrap samples

This chapter splits up into two parts: tuning parameter selection through information criteria on bootstrap samples is studied in the first part, while the second part deals with tuning parameter selection through cross-validation performed on bootstrap samples. In both parts simulations and real data studies are performed to investigate and compare the prediction accuracy and the complexity of gradient boosting models fit on bootstrap samples to those of models fit on original samples. Promising alternatives, such as subsampling, are also investigated.

Publications and contributing manuscripts

Large parts of this dissertation are based on publications where I am the main contributor. The works were supervised by Anne-Laure Boulesteix and were prepared in cooperation with Gerhard Tutz from the Department of Statistics of the University of Munich, Ender Celik who is a graduate student of biostatistics and Harald Binder from the University Medical Center Mainz. The works are named in the following:

- Janitza S., Tutz G. and Boulesteix A.-L. (2016). Random forest for ordinal responses: Prediction and variable selection. Journal of Computational Statistics & Data Analysis 96:57–73. (Chapter 3)

- Janitza S., Celik E. and Boulesteix A.-L. (2015). A computationally fast variable importance test for random forests for high-dimensional data. Technical Report 185, University of Munich. (Chapter 4)

For this paper I received the "Student/Postdoctoral Fellow Paper Competition and Travel Award 2015" by the IFCS and I was invited to submit the paper for possible publication in Advances in Data Analysis and Classification.

- Janitza S., Binder H. and Boulesteix A.-L. (2016). Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications. Biometrical Journal 58(3):447–473. (Chapter 5 and the first part of Chapter 6)

I also contributed to further publications on random forests or other resampling approaches. In Chapters 2 and 5 I refer to these works and give a brief summary of some of their results. The publications are outlined in the following.

- Dolch M.E.*, Janitza S.*, Boulesteix A.-L., Grassmann C., Praun S., Denzer W., Schelling G. and Schubert S. (2016). Gram-negative and -positive bacteria differentiation in blood culture samples by headspace volatile compound analysis. Journal of Biological Research-Thessaloniki 23:3. (* joint first co-authorship)
  This practical work was in close collaboration with medical doctors from the University Hospital Munich. A random forests prediction rule was derived which shows good differentiation between gram-negative and -positive bacteria.

- Boulesteix A.-L., Janitza S., Kruppa J. and König I.R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(6):49–507.
  This review paper on random forests was in collaboration with Inke König and Jochen Kruppa from the University Hospital Lübeck.

- Boulesteix A.-L., Janitza S., Hapfelmeier A., van Steen K. and Strobl C. (2015). Letter to the Editor: On the term "interaction" and related phrases in the literature on random forests. Briefings in Bioinformatics 16(2):338–345.
  This letter to the Editor was written in response to the random forests review of Touw et al. (2012), an often cited article recently published in the high-impact journal Briefings in Bioinformatics. The letter aims at the clarification of the imprecise statements made by Touw et al. (2012).

- Rospleszcz S.*, Janitza S.* and Boulesteix A.-L. (2016). Categorical variables with many categories are preferentially selected in model selection procedures for multivariable regression models on bootstrap samples. Biometrical Journal 58(3):652–673. (* joint first co-authorship).

This work is the result of a master thesis which I supervised. It addresses a specific problem related to the use of the bootstrap in model building procedures.

- De Bin R., Janitza S., Sauerbrei W. and Boulesteix A.-L. (2016). Subsampling versus bootstrap in resampling-based model selection for multivariable regression. Biometrics 72(1):272–280.
  This work resulted from a collaboration of our working group with Willi Sauerbrei from the University Medical Center Freiburg. It compares subsampling with bootstrapping for the use in model building procedures.

Software

All computations were carried out using the statistical software R (R Core Team; 2013) and related packages. Chapters 3 and 4 are mainly based on the libraries party (Hothorn, Hornik and Zeileis; 2006) and randomForest (Liaw and Wiener; 2002), respectively. Further packages are indicated in the respective chapters. The new variable importance measures for ordinal responses (Chapter 3) were implemented in the statistical software R and are available from the website `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/janitza/rf_ordinal/index.html`. These functions make use of the library party (Hothorn, Hornik and Zeileis; 2006). The new computationally fast variable importance test (Chapter 4) is implemented in the R package vita (written by Ender Celik).

2. Trees, bagging and random forests

The first part of this chapter introduces the concept of decision trees which build the basis for bagging and the ensemble method random forests (RF). Different tree building algorithms exist, such as classification and regression trees, in brief CART (Breiman et al.; 1984), conditional inference trees (Hothorn, Hornik and Zeileis; 2006), C4.5 (Quinlan; 1986) and ID3 (Quinlan; 1993), to name just a few (see Loh; 2014, for a recent overview). In this chapter, two versions of binary decision trees are described, namely CART that are used in the original RF version of Breiman (2001) and conditional inference trees that are used in the RF version of Hothorn, Hornik and Zeileis (2006). In the second part of this chapter, a brief review of bagging and RF is given. A special emphasis is put on RF's out-of-bag based variable importance measures since some modifications of these measures are presented in Chapters 3 and 4. The chapter closes with an application example of RF in medicine described in Dolch et al. (2016). This paper is a result of a close cooperation with Michael Dolch and his colleagues from the Department of Anesthesiology of the University Hospital Munich.

2.1. Trees

Decision trees are a non-parametric method for predicting the response Y from the predictor variables $X_1, \ldots, X_p$. If the response variable is metric, the trees are termed *regression trees*, and in the case of a categorical response one speaks of *classification trees*. The idea of decision tree methodology consists in recursively partitioning the data into subsets which are more homogeneous with respect to the response variable. The recursive splitting process is illustrated in the tree diagram in Figure 2.1. The root node of the decision tree contains all observations $i = 1, \ldots, n$. Starting at the root node, the observations are recursively partitioned into two daughter nodes based on their predictor values $x_{i1}, \ldots, x_{ip}$. In the tree diagram in Figure 2.1 there are two metric or ordinal variables, X_1 and X_2 (i.e., $p = 2$). The first split is implemented using variable X_1. The root node is partitioned into two disjoint subsets $\{i|x_{i1} \leq c_1\}$ and $\{i|x_{i1} > c_1\}$ based on the variable X_1 and the cutpoint c_1. Note that if X_1 was a categorical variable with levels $\{1, \ldots, m\}$ without any natural order, the partition would result in the subsets $\{i|x_{i1} \in S_1\}$ and $\{i|x_{i1} \notin S_1\}$, with

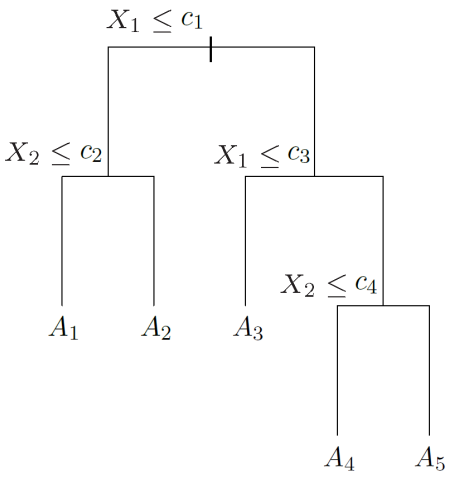

Figure 2.1.: Illustration of a decision tree in which the data is recursively partitioned based on the variables X_1 and X_2.

$S_1 \subset \{1, \ldots, m\}$. The resulting subsets are referred to as the *daughter nodes* which are then again split into two daughter nodes. Let us consider as an example the daughter node containing the observations $\{i|x_{i1} \leq c_1\}$. The observations of this daughter node are further partitioned using the variable X_2 and the cutpoint c_2. This results in daughter nodes containing the observations $\{i|x_{i1} \leq c_1 \wedge x_{i2} \leq c_2\}$ and $\{i|x_{i1} \leq c_1 \wedge x_{i2} > c_2\}$, respectively. The splitting process is continued in each daughter node until either the daughter node cannot be split any more or a stopping criterion is fulfilled. The nodes of the tree that are not partitioned further are termed *terminal nodes* or *leaves*. In Figure 2.1 the terminal nodes correspond to the subsets $A_1, \ldots, A_5$. All other nodes of the tree are referred to as *inner* or *internal nodes*.

The tree prediction for a new observation is obtained by tracking the observation down the tree until it arrives at a terminal node. The prediction for the new observation is established based on the observations contained in the terminal node. In the case of a regression tree, the prediction usually corresponds to the mean response value. The prediction for a new observation falling into the terminal node A_l is then given by

$$\hat{Y} = \frac{1}{|A_l|} \sum_{i \in A_l} y_i, \tag{2.1}$$

with $|A_l|$ denoting the cardinality of A_l.

In the case of classification trees, the terminal nodes predict probabilities for the response classes $r = 1, \ldots, k$. The probability for class r which is predicted for a new observation falling into the terminal node A_l is defined as

$$\hat{\pi}_r = \frac{1}{|A_l|} \sum_{i \in A_l} I(y_i = r), \tag{2.2}$$

where $I(\cdot)$ denotes the indicator function. The predicted class can then be obtained as the class for which the predicted probability is maximal:

$$\hat{Y} = \underset{r=1,\ldots,k}{\arg\max}\, \hat{\pi}_r. \tag{2.3}$$

The selection of variables and cutpoints for splitting the nodes in a tree are based on certain splitting criteria. Two specific split selection strategies used in CART and in conditional inference trees are outlined in the following two sections.

2.1.1. Split selection in classification and regression trees

In classification and regression trees (CART), splits are implemented based on so-called *impurity measures*. For classification trees impurity refers to the distribution of response classes observed for the observations contained in a node. If many of the observations in a node have the same response class, the node is considered pure. It is purest when all observations have the same response class. In contrast to that, when a node contains an equal amount of observations from all classes, the node impurity is largest. There are different measures which are used to quantify node impurity. The most popular measures for classification trees are the *Gini index* and the *Shannon entropy*. The empirical version of the Gini index for a node A is defined as

$$I_G(A) = 1 - \sum_{i=1}^{k} \left(\frac{n_i(A)}{n(A)} \right)^2,$$

where $n_r(A)$ is the number of observations from class $r \in \{1, \ldots, k\}$ that are contained in the node A and $n(A)$ denotes the total number of observations in the node A. The Gini index is smallest and takes value 0 in the case where one of the class frequencies $n_1(A)/n(A), n_2(A)/n(A), \ldots, n_k(A)/n(A)$ equals 1. In this case, all observations in the node belong to the same class. Thus the smaller the Gini index, the purer the node, and the node is purest if the Gini index is 0. The Gini index is maximal for $n_1(A)/n(A) = n_2(A)/n(A) = \ldots = n_k(A)/n(A)$, that is, all classes are represented by an equal number of observations in the node.

An alternative impurity measure is the Shannon entropy. Its empirical version for a node A is given by

$$I_E(A) = -\sum_{i=1}^{k} \frac{n_i(A)}{n(A)} \log_2 \left(\frac{n_i(A)}{n(A)} \right).$$

The Shannon entropy is 0 if all observations in the node belong to the same class and has larger values with an increasing balance between the classes.

In regression trees the nodes with smaller variability in the response values are consid-

ered more pure. A natural measure for node impurity in regression trees is thus the mean squared error.

When constructing trees the impurity of a node is reduced by splitting the node into two more homogeneous daughter nodes. The variable and the cutpoint that yield the maximal reduction in node impurity according to a pre-defined impurity measure are chosen for the split. If $I_M(A)$ denotes the impurity of node A which is measured through an arbitrary impurity measure M, and $I_M(A_{left})$ and $I_M(A_{right})$ denote the impurities of the left and right daughter nodes, the decrease in node impurity by splitting the node A into the two daughter nodes A_{left} and A_{right} is given by

$$\Delta_M(A|A_{left}, A_{right}) = I_M(A) - \left(\frac{n(A_{left})}{n(A)} I_M(A_{left}) + \frac{n(A_{right})}{n(A)} I_M(A_{right}) \right).$$

In CART the split among all possible splits is selected for which the decrease in node impurity is largest. This involves a search through all variables and through all possible split points of a variable. Note that some variables offer more split points than others. A nominal variable with m levels offers $2^{m-1} - 1$ possible splits, while a metric variable with n distinct values has $n - 1$ possible splits, and an ordinal variable with m levels has only $m - 1$ possible splits. Thus when variables of different scales are considered, there is a preferential selection of variables which have many possible splits because the chance that the optimal split is found in these variables is higher. For example, a nominal variable with many categories is preferentially selected for a split over nominal categorical variables with fewer categories or over metric or ordinal variables.

Moreover, there is a preferential selection of categorical variables with balanced categories over categorical variables with unbalanced categories (Nicodemus; 2011; Boulesteix, Bender, Bermejo and Strobl; 2012), and variables with many missing values are preferentially selected over variables with fewer or no missing values (Kim and Loh; 2001). Alternative strategies for selecting the optimal split variable and cutpoint which prevent these issues are based on hypothesis testing. There are some tree methods which make use of hypothesis tests for split selection (Loh and Shih; 1997; Kim and Loh; 2001). A specific tree methodology which makes use of permutation testing procedures is outlined in the following section.

2.1.2. Split selection in conditional inference trees

In the conditional inference trees of Hothorn, Hornik and Zeileis (2006), conditional inference tests are performed for selecting the best split in an unbiased way. In contrast to CART, the variable selection is separated from the split point selection when selecting the optimal split. For each split in a tree, each variable is tested for its association with

the response, yielding a p-value. The variable with the smallest p-value is selected for a split. In the next step the best split point within the variable is chosen. The selection of a split variable based on its p-value has the advantage that there is no preference for certain types of variables.

The algorithm of Hothorn, Hornik and Zeileis (2006) implements the following steps which are recursively repeated in all daughter nodes:

1. *Stopping criterion*: Test the global null hypothesis which states that none of the variables is associated with the response. If the p-value falls below a prespecified significance level α, the global null hypothesis of independence is rejected and the next step is performed, otherwise the recursion is stopped.

2. *Split variable selection*: Perform a test of independence between each variable and the response and select the variable X_{j*} with the smallest p-value to implement the split.

3. *Split point selection*: Perform a special two-sample-test for all possible binary split points within the variable X_{j*} and choose the split point that yields the smallest p-value.

The methodology of Hothorn, Hornik and Zeileis (2006) utilizes a permutation test framework and is thus applicable to problems where both predictors and response are measured on arbitrary scales, including nominal, ordinal, discrete and continuous variables. Moreover, the methodology even applies to multivariate responses.

Steps 1 and 2 make use of the same test statistics computed for tests of independence between each of the variables and the response. In step 1 the p-values corresponding to the test statistics are adjusted for multiple testing and are subsequently compared to the significance level α. The computation of the linear statistics which are the basis for the test statistics from which p-values are derived, is outlined in the following.

The statistic that is used for testing the association between the response Y and a predictor variable X_j based on observations $i = 1, \ldots, n$ is defined as

$$T_j = \text{vec}\left(\sum_{i=1}^{n} g_j(X_{ij}) h(Y_i, (Y_1, \ldots, Y_n))^\top\right) \tag{2.4}$$

with $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ being a non-random transformation of the predictor variable, and $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ being a function that depends on the response vector $(Y_1, \ldots, Y_n)^\top$ in a permutation symmetric way [1]. Permutation symmetry means that h does only depend on

[1] In contrast to the statistic given in Hothorn, Hornik and Zeileis (2006), all observations $i = 1, \ldots, n$ are used for deriving the test statistic (thus omitting observation weights).

the order statistics $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$, with $Y_{(j)}$ being the j-th smallest value of the sample, but not on the sequential order of the data (see also Strasser; 2000).

The specification of the functions h and g_j depends on the scale of the response and predictor variables, respectively (see also Hothorn, Hornik and Zeileis; 2006, for examples). For a nominal response taking levels in $1, \ldots, q$, the function h transforms the response to the unit vector of length q. For the computation of the statistic, this vector is multiplied with $g_j(X_{ij})$ which is a p_j-dimensional vector – the result being a matrix of dimension $p_j \times q$. The vec operator then converts the $p_j \times q$ matrix into a $p_j q$ column vector, so that the statistic T_j is itself an $p_j q$-dimensional vector. The statistic is then mapped onto the real line, for example by taking the component that has maximal absolute standardized value; see Hothorn, Hornik and Zeileis (2006). For numeric responses h might simply be chosen as the identity function such that $h(Y_i, (Y_1, \ldots, Y_n)) = Y_i$ and $q = 1$.

The specification of the functions g_j is similar to the specification of h. For a numeric predictor variable the transformation is usually the identity function such that $g_j(X_{ij}) = X_{ij}$ and $p_j = 1$. For a nominal predictor variable taking levels in $1, \ldots, m$, g_j transforms the predictor variable to the unit vector of length m and $p_j = m$. For an ordinal predictor variable the levels are transformed to a metric scale through attributing scores to the levels of the variable. The transformed variable is then handled in the same way as a metric predictor variable.

In the case of an ordinal response the response is transformed to a metric scale by attributing scores – but now scores are attributed to the levels of the response variable. If $s(r) \in \mathbb{R}$ denotes the score for category $r \in \{1, 2, \ldots, k\}$ and Y_i denotes the ordinal response of observation i with covariates $X_{ij}, j = 1, \ldots, p$, the statistic simplifies to

$$T_j = \sum_{i=1}^{n} g_j(X_{ij}) s(Y_i). \tag{2.5}$$

Trees that are constructed based on the statistic (2.5) are denoted by *ordinal regression trees*.

Note that the test statistic for an ordinal response coincides with a test statistic for a numeric response with values $s(Y_1), \ldots, s(Y_n)$. This leads to the selection of the same variables and cutpoints in ordinal regression trees and regression trees. Though ordinal regression trees and regression trees have the same tree structure, predictions by the trees are different. Ordinal regression trees provide predicted classes or estimates for class probabilities (cf. Eq. (2.3) and (2.2)), while regression trees give real-valued predictions (cf. Eq. (2.1)). The tree construction of ordinal regression trees thus corresponds to that of regression trees, while predictions are obtained in the same way as for classification trees.

The statistics of the form (2.4) are standardized to obtain test statistics. From the test statistics p-values can be derived (see Hothorn, Hornik, Van De Wiel and Zeileis; 2006;

Hothorn, Hornik and Zeileis; 2006, for detailed information). The variable with the smallest p-value is then selected for splitting a node in step 2 of the algorithm (cf. p. 13).

For implementing a split in step 3 the optimal cutpoint within the selected variable has to be chosen. All possible cutpoints within the variable which partition the data into two subsets are considered. In conditional inference trees the cutpoint is chosen that maximizes the discrepancy in the response values between the two groups of observations which are defined by the binary split. For each possible split of the sample space S into the disjoint subsets S_1 and S_2, with $S = S_1 \cup S_2$, a two-sample statistic is used which measures the discrepancy in the response between observations $\{i|x_{ij} \in S_1\}$ and observations $\{i|x_{ij} \in S_2\}$. This two-sample statistic emerges as a special case of Eq. (2.4), in which the function g_j is the indicator function which takes value 1 if x_{ij} is contained in S_1, and 0 otherwise. Among all possible splits the split which maximizes the two-sample test statistic is chosen.

Note that the term "conditional" in the name "conditional inference trees" refers to the property of the tests used as splitting criterion for split selection. This has led to confusion in the literature; for example, Touw et al. (2012) incorrectly lead the term "conditional" in the term conditional inference trees back to the conditional variable importance proposed by Strobl et al. (2008) and to interactions between variables. A clarification was given by Boulesteix et al. (2015) in a letter to the Editor.

2.2. Bagging and forests

2.2.1. Bootstrap and bootstrap aggregating ("Bagging")

The bootstrap method proposed by Efron and Tibshirani (1993) has become a popular tool that is applied in diverse areas. In brief, the main idea of the bootstrap procedure is that the sample is treated as the population and the estimates of the sample are treated as the true population parameters. Instead of sampling from the true distribution, with the non-parametric bootstrap (considered in this thesis) one randomly draws n observations with replacement from the observed data $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$. The resulting bootstrap sample $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)^\top$ contains as many observations as the original sample. By drawing from the original sample with replacement, some observations are contained several times in the bootstrap sample while other observations are not at all contained in the sample. If n is chosen large enough the probability that an observation is not contained in the bootstrap sample can be approximated as

$$\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx 0.368.$$

Hence 63.2% of the observations from the original data are contained in a bootstrap sample at least once. If we observe pairs of observations $z_i := (y_i, x_i^\top)^\top$, with $x_i = (x_{i1}, \ldots, x_{ip})^\top$ as the covariate vector and y_i as the response for observations $i = 1, \ldots, n$, the covariates and the response are sampled together, that is, the bootstrap sample $z^* = (z_1^*, z_2^*, \ldots, z_n^*)^\top$ is obtained.

Bootstrapping involves drawing a large number B of bootstrap samples from the original data. It is used for different purposes, usually to derive standard errors or confidence intervals when it is difficult or problematic to derive these based on asymptotic theory. Further, it is often used for the estimation of the error of a prediction modeling strategy. This is also addressed in Chapter 6. An explanation of the various applications of the bootstrap is beyond the scope of this thesis. The interested reader is referred to the book by Efron and Tibshirani (1993) which gives an extensive overview. In this chapter, the focus is laid on **b**ootstrap **agg**rega**ting**, in brief bagging. Bagging is a special application of the bootstrap that aims at improving the prediction performance of a learning algorithm. It was developed by Breiman (1996a) and combines the predictions which are obtained from prediction rules constructed on bootstrap samples. More precisely, B prediction rules are constructed, with the b-th prediction rule being based on bootstrap sample z^{*b}. If the response is numeric, the prediction of the bagged predictor is obtained by averaging the predicted values over the B prediction rules. If the response is categorical, the bagged predictor predicts the class which is most often predicted by the B prediction rules. This is often referred to as the *majority vote* or *plurality vote*.

Depending on the considered method, the prediction rules may also give probabilities for the response classes. Two different approaches for obtaining the class predictions of a bagged predictor were considered by Breiman (1996a). One approach transforms the predicted class probabilities that are obtained by a prediction rule into predicted classes by using the class with the largest predicted probability. Then the aggregated prediction of the bagged predictors is obtained through majority voting. The other approach aggregates the class probabilities over the B prediction rules and then decides for the class with the largest averaged probability. If $\hat{\pi}_r^b$ denotes the probability of an observation belonging to class r that is predicted by the b-th prediction rule, the prediction by the bagged predictor is

$$\hat{Y} = \underset{l=1,\ldots,k}{\arg\max} \left\{ \frac{1}{B} \sum_{b=1}^{B} \hat{\pi}_l^b \right\}. \tag{2.6}$$

In the studies of Breiman (1996a) the misclassification rate of the bagged predictor was almost identical for both approaches.

Through aggregation, bagging considerably reduces the variance of unstable procedures without changing the bias (see, e.g., Hastie et al.; 2001, for a proof). This results in

prediction accuracies of bagged predictors which are better than those of single prediction models, as shown by several authors (Kohavi and Kunz; 1997; Bauer and Kohavi; 1999; Maclin and Opitz; 1997; Dietterich; 2000). Trees have often been used as base learners in the bagging procedure. The bagged predictor reduces the large variance exhibited by single decision trees while benefiting from their low bias. However, the better accuracy comes at a cost of the good interpretability of trees which is lost in the bagging process. The accuracy of more stable procedures such as nearest neighbors, in contrast, was shown to be hardly improved by bagging (Hastie et al.; 2001).

As will be seen in the next section, the RF method which is nowadays widely applied is rooted in the bagging procedure.

2.2.2. Random forests

The random forest (RF) method (Breiman; 2001) is a modification of bagging trees, that enforces diversity between base classifiers. In bagged trees each tree is build based on a different sample randomly drawn with replacement from the data. It is clear that bagged trees differ in their structure because each tree is based on a different random sample of the observations. In addition to this, RF further encourages diversity by considering only a subset of the variables for implementing a split, whereby for each split in a tree the subset is randomly drawn from all p variables. This random subset selection leads to trees that are more different to each other than bagged trees are. Breiman (2001) shows that by introducing diversity between trees, the resulting ensemble might have better prediction performance as the ensemble accuracy depends on two factors: the accuracy of the single trees and the dependence between the trees.

The number of randomly drawn variables that are considered for a split is referred to as *mtry* in the RF literature. Which values for *mtry* are appropriate depends on the considered problem. The parameter *mtry* should therefore ideally be tuned in practical applications. Extreme values for *mtry*, such as *mtry* close 1 or p, are, however, not recommended. With an *mtry* value chosen too small, trees are more diverse but important variables might not be selected for splitting and the resulting RF might have poor predictive power. If *mtry* is chosen too large, predictors with strong effects are often selected for a split but it hinders the selection of predictors with small effects. In addition, large values for *mtry* lead to trees that are less diverse. In the extreme case in which *mtry* is chosen as the total number of predictors p, RF corresponds to bagging trees.

Like bagging, RF do not suffer from overfitting and thus trees in RF do not have to be pruned (Breiman; 2001). That is why trees are grown to full size in all methodological studies presented in this thesis. In practical applications, trees might be pruned in order to optimize prediction accuracy. In this case, parameters have to be tuned that control

the size of trees, such as the minimal number of observations that are required in a node. Parameter tuning for controlling tree size was also performed in the medical application example at the end of this chapter. More information on tuning parameters that control the size of trees and recommendations about their choice are given, for example, in Boulesteix, Janitza, Kruppa and König (2012).

The aggregation of individual classification tree predictions differs for the two RF versions considered in this thesis. The classical RF version of Breiman (2001) uses the trees' predicted classes and derives the ensemble prediction as the majority vote, that is the class which is predicted most often by the trees. In contrast to that, the RF version of Hothorn, Hornik and Zeileis (2006) uses the predicted class probabilities and derives the ensemble prediction based on the average of class probabilities (Eq. (2.6)). In contrast to many other prediction methods, RF allows using the same data for constructing the RF and evaluating its prediction performance. This is outlined in the following.

Out-of-bag observations

Since each tree is built from a random sample of the data, there are some observations in the data which were not used in its construction. These observations are denoted by *out-of-bag (OOB) observations*. If bootstrap samples are used to construct trees, about 63.2% of the observations from the original data are contained in the bootstrap sample at least once. It follows that approximately 36.8% of the observations are not contained in the sample and are thus OOB observations (see Section 2.2.1). If subsampling is used, the user specifies the proportion of OOB observations. The OOB observations are often used for assessing the prediction performance of a RF. The idea is to derive the prediction for an observation from only those trees which were not constructed based on this observation (i.e., the trees for which the observation is OOB). In this way, the predictions for all observations are obtained, and one can estimate the RF's prediction error using, for example, the error rate in the case of a categorical response, or the mean squared error in the case of a metric response. Since the error is computed based on the OOB observations, it is then referred to as the *OOB error* (Breiman; 1996b). The OOB error is often considered as a good estimator of the error which is expected for independent data, although some studies suggest that this might not be the case (Mitchell; 2011; Bylander; 2002).

The OOB observations have not only proven useful for estimating the prediction error of a RF but also for computing the RF's permutation variable importance as outlined in the following section.

Variable importance measures

RF provides measures that can be used for obtaining a ranking of predictors. The ranking reflects the importance of these variables in the prediction of the response and it might be used to select the variables with the best predictive ability. The two standard variable importance measures (VIMs) implemented in the RF version of Breiman (2001) are the permutation VIM (also referred to as the mean decrease in accuracy) and the Gini VIM. The latter prefers certain types of predictors (Strobl et al.; 2007; Nicodemus and Malley; 2009; Nicodemus; 2011; Boulesteix, Bender, Bermejo and Strobl; 2012) and therefore its predictor rankings should be treated with caution (see Boulesteix, Janitza, Kruppa and König; 2012, for an overview). This thesis focuses on the permutation VIM which gives essentially unbiased predictor rankings. A general definition of the permutation VIM is used, in which the trees' prediction error is measured by an arbitrary error measure M (e.g., the error rate).

According to the permutation VIM, the variable importance of variable X_j is then

$$VI_j^M = \frac{1}{ntree} \sum_{t=1}^{ntree} (MP_{tj} - M_{tj}), \tag{2.7}$$

where

- $ntree$ denotes the number of trees in the forest,
- M_{tj} denotes the error of tree t when obtaining predictions for all OOB observations *before* permuting the values of predictor variable X_j, and
- MP_{tj} denotes the error of tree t when obtaining predictions for all OOB observations *after* randomly permuting the values of predictor variable X_j.

The idea underlying this VIM is the following: if the predictor is not associated with the response, the permutation of its values has no influence on the classification, and thus no influence on the trees' performance. Then the error of the trees is not substantially affected by the permutation and the importance score of the predictor takes a value close to zero, indicating that there is no association between the predictor and the response. In contrast, if response and predictor are associated, the permutation of the predictor's values destroys this association. "Knocking out" this predictor by permuting its values results in worse prediction. The difference in errors after and before randomly permuting the predictor takes a positive value, reflecting the high importance of this predictor.

The two established permutation VIMs for RF arise when using the error rate (for classification trees) or the mean squared error (for regression trees) as the performance measure M in Eq. (2.7). Throughout this thesis these measures will be termed the *error rate*

based (permutation) VIM and the *MSE-based (permutation) VIM*, respectively. These VIMs have been explored in the literature in the context of classification and regression tasks, respectively, and are often applied in the literature (e.g., Steidl et al.; 2010; Karamanian et al.; 2014; Harrington et al.; 2014).

Unsolved problems

RF are often claimed to incorporate complex interactions between predictors. This makes RF especially attractive for high-dimensional complex genetic data. There is no question that the structure of classification and regression trees can advantageously take interaction effects into account. If a node A is split by predictor variable X_1 into the daughter nodes A_{left} and A_{right}, then the effect of another predictor variable X_2 may be substantially different in the two daughter nodes. This would indicate that there is an interaction effect between X_1 and X_2. In contrast to that, if there was no interaction effect, then one would ideally expect that the same split variable and the same split point is chosen for the two child nodes (Boulesteix et al.; 2015). However this situation rarely occurs in practice. Firstly because there are random variations in the data, which is especially pronounced in small samples. And since the data is recursively split, the number of observations in the nodes gradually decreases until there are only few observations in each leaf; so even for data with large observation numbers one cannot rule out that random variations lead to the selection of different splits. Secondly, in RF the splitting variable is selected from a subset of the variables. The subset of *mtry* variables which is considered for a split is drawn anew for each node, and thus it can happen that a variable X_2 which was selected for node A_{left} is not among the *mtry* variables for the node A_{right}. Due to these two reasons it is extremely rare that a tree selects the same predictor variable and the same split points for both nodes. Thus, in practice a tree almost always looks as if there were interactions. But such patterns will also be seen in the absence of interactions due to the reasons mentioned above. The question is thus whether these patterns are just the result of random variations (with respect to the sample or with respect to the set of *mtry* candidate predictors) or of true interactions. This question is far from trivial and to date there exists no standard approach to answering it from looking at the trees' structure (Boulesteix et al.; 2015).

An application example

The occurrence of infectious complications in critically ill patients greatly affects patient outcome and leads to increased mortality rates. A fast identification of the causative organism is of highest priority as this allows the early administration of antibiotic treatment. The first step in this process is the detection of microorganism growth in blood cul-

ture broth bottles. This has achieved high reliability and works on a semi-automatic basis. However, the subsequent process of microorganism identification is time-consuming and requires staff presence. There is an urgent need of rapid and reliable diagnostic methods facilitating the identification of the causative microorganisms. Dolch et al. (2016) applied ion-molecule reaction mass spectrometry analysis of headspace gas volatile compound composition to differentiate between microorganisms by using blood culture broth samples. An RF classifier was found which achieved high accuracy in differentiating between Gram-positive and Gram-negative bacteria and might be promising for rapid diagnosis which allows for prompt antibiotic treatment.

In the studies several prediction methods were considered for differentiating between Gram-positive and Gram-negative bacteria, such as boosting, RF, support vector machines, penalized logistic regression, k nearest neighbors, feed forward and probabilistic neural networks, discriminant analysis, elastic net and lasso-type methods. Where applicable, methods were also applied after variable selection and/or dimension reduction. A complete list of the considered classifiers is given in Appendix A. To reliably assess the performance of the prediction rules and avoid over-optimism, a random split validation procedure was adopted (Daumer et al.; 2008; Boulesteix and Strobl; 2009). This means that prior to analysis the data was randomly split into two non-overlapping sets (ratio 2:1). Stratified splitting was conducted to preserve the distribution of Gram-positive and -negative bacteria in both sets. The larger set (denoted by training set) was used for training the candidate prediction rules and for selecting the best one. The smaller set (denoted by validation set) was used for the validation of the best prediction rule in order to obtain a reliable estimate of the expected performance of the prediction rule on future independent data.

The error rate of the candidate prediction rules in the training set was assessed based on 5-fold cross-validation. To obtain more stable estimates for the error rate 100 repetitions of 5-fold cross-validation were conducted. The prediction rule with the smallest cross-validated error rate was regarded as the best. A RF classifier based on the 10 variables with the highest Gini variable importance achieved the smallest cross-validated error rate (9.1%), and was thus considered the best prediction rule (see Appendix A for details on the computations and further results). Majority vote was used to classify bacteria as either Gram-positive or Gram-negative. The portion of Gram-positive bacteria correctly classified as Gram-positive (defined as sensitivity) was 97.5% in the training set. The portion of Gram-negative bacteria correctly classified as Gram-negative (specificity) was 74.8%, and the area under the curve was 0.93. The higher portion of correctly classified Gram-positive bacteria is likely to be due to the tendency of RF for predicting the larger class (i.e., the Gram-positive bacteria) (see Janitza et al.; 2013, and references therein).

	Training set ($n = 86$)	Validation set ($n = 42$)
Error rate	0.091	0.167
Sensitivity	0.975	0.933
Specificity	0.748	0.583
Area under the curve	0.93	0.89

Table 2.1.: Performance of the RF prediction rule on the training and validation sets.

The RF prediction rule was applied to the validation set and achieved good performance also for the validation set (see Table 2.1). The performance measured by the error rate, the frequency of Gram-positive bacteria correctly classified (sensitivity), the frequency of Gram-negative bacteria correctly classified (specificity) and the area under the curve, was better on the training set than on the validation set. Such a result is also obtained in settings where no associations between the predictor variables and the response exist if the best prediction rule is chosen out of a large number of candidate prediction rules (Boulesteix and Strobl; 2009). One can, however, assume that the good performance of the RF classifier on the training set is not a result of an optimization procedure since the RF prediction rule showed good performance also on the validation set, with an error rate of 16.7% and an area under the curve of 0.89. Figure 2.2 shows the ROC curve for the RF prediction rule applied to the validation set. The practical utility of the RF prediction rule as a diagnostic tool has to be assessed by clinicians.

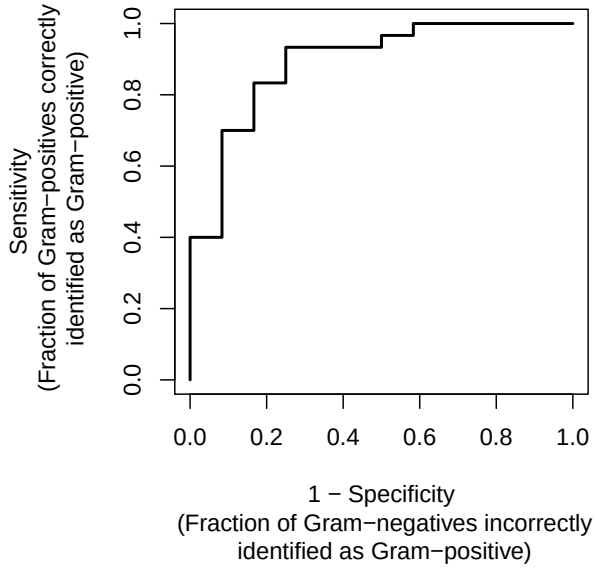

Figure 2.2.: ROC curve for the RF prediction rule evaluated on the validation set.

3. Random forests for ordinal responses: prediction and variable selection

This chapter is based on Janitza, Tutz and Boulesteix (2016). It investigates the use of RF for prediction and variable selection purposes in cases where the response has an inherent ordering. After a general introduction the methods are introduced in Section 3.2. The first part of the methods section reviews established performance measures that can be used to assess the ability of a classifier to predict an ordinal response. The second part outlines novel VIMs that are proposed for predictor rankings through RF and ordinal response data. In Sections 3.3 and 3.4 studies on simulated and real data, respectively, are presented. In both sections the studies of prediction performance are first reported. Here the prediction performance of a RF constructed from classification trees is compared with that of a RF constructed from ordinal regression trees. Subsequently the studies on VIM performance are shown in which the performance of the standard error rate based VIM is compared to those of the three alternative permutation VIMs when computed on classification and ordinal regression trees. In Section 3.5 the findings are summarized and recommendations to applied researchers working with RF and ordinal response data are given.

3.1. Introduction

In many applications where the aim is to predict the response or to identify important predictors, the response has an inherent ordering. Examples of ordinal responses in biometrical applications are tumor stages I - IV, disease severity, for example from mild to moderate to severe disease state, and artificially created scores combining several single measurements into one summary measure, like the Apgar score, which is used to assess the health of a newborn child. Appropriate handling of ordinal response data for class prediction as well as for feature selection is essential to efficiently exploit the information in the data. A study concerning stroke prevention showed that statistical efficiency

was much higher when using an ordinal response such as fatal/nonfatal/no stroke compared to a binary outcome providing only the information of whether a patient had a stroke or not (Bath et al.; 2008). Statistical models for ordinal response data such as the proportional odds, the continuation ratio and the adjacent category model have been investigated extensively in the literature (see Agresti; 2002). However, these methods are not suitable for applications where the association between predictors and the response is of a complex nature, including higher-order interactions and correlations between predictors. Moreover, the models rely on assumptions (such as proportional odds) that are frequently not realistic in practical applications. Further, parameter estimation typically faces the problem of numerical instability if the number of predictors is high compared to the number of observations.

For nominal and numeric response the application of RF has been well investigated. However, in the case of ordinal response there is no standard procedure and literature is scarce. While in the classical RF algorithm by Breiman (2001) the ordering of a predictor is taken into account by allowing splits only between adjacent categories, the ordering information in the response is ignored (i.e., the response is treated as a nominal variable), and an ensemble of classification trees is constructed. However, ignoring the ordering information results in a loss of information. For single classification and regression trees (CART) several approaches for predicting an ordinal response have been developed. These are based on alternative impurity measures to the Gini index. Prominent examples are the ordinal impurity function suggested by Piccarreta (2001) and the generalized Gini criterion introduced by Breiman et al. (1984). With these measures a higher penalty is put on misclassification into a category that is more distant to the true class than on misclassification into a category that is close to the true class, thus taking into account the ordinal nature of the response. The ordered twoing criterion by Breiman et al. (1984, p. 38) is another popular measure that does not rely on misclassification costs but rather on reducing the k-class classification problem to $k-1$ two-class classification problems where a split that divides the k classes into two classes is only made between adjacent categories (see Breiman et al.; 1984, for a detailed description). Archer and Mas (2009) investigated the prediction accuracy of bagged trees constructed through the ordered twoing method (Breiman et al.; 1984) and the ordinal impurity function (Piccarreta; 2001) for classifying an ordinal response. Using simulation studies they showed that the ordered twoing method and the ordinal impurity function are reasonable alternatives to the Gini index in tree construction. However, in their real data application these measures did not perform better than the Gini index. Except for the study of Archer and Mas (2009), approaches for ordinal regression problems have only been discussed for CART and have not been extended to RF.

The unbiased RF version of Hothorn, Hornik and Zeileis (2006) is a promising tool for constructing trees with ordinal response data because, in contrast to the standard RF implementation by Breiman (2001), where splitting is based on the Gini index, it provides the possibility of taking the ordering information into account when constructing a tree. The resulting trees are denoted by *ordinal regression trees*. For constructing ordinal regression trees one has to attach scores to each category of the ordinal response. These scores reflect the distances between the levels of the response. When the response is derived from an underlying continuous variable, the scores can be chosen as the midpoints of the intervals defining the levels. For example, when creating categories for different smoking levels, Mantel (1963) suggested defining the scores as the average number of cigarettes per day or week. Note that when defining scores only the relative spacing of the scores is important, not the absolute; for example the scores 1, 2, 3 reflect the same relative distance between categories as the scores 1, 3, 5. The advantage of using the ordering of a variable is that tests which take the ordering into account have higher power compared to tests which ignore the underlying ordering because some degrees of freedom are saved by restricting the possible parameter space (Agresti; 2002, p. 98).

A further issue which is investigated is the appropriate handling of the ordering information in the response when computing VIMs. The importance score for each predictor is derived from the difference in prediction performance of the single trees resulting from the random permutation of this predictor. For numeric responses the mean squared error of the predicted and the true values is used as the prediction performance measure to compute the importance. For categorical responses (nominal and ordinal) the standard is to use the error rate. An appropriate prediction performance measure is essential for a good VIM performance, as demonstrated by Janitza et al. (2013), who showed that in the case of two response classes which differ in their class sizes the area under the curve is a more appropriate performance measure for computing the importance score of a predictor than the commonly used error rate.

The design of an appropriate VIM in the common case of ordinal response variables, however, has not been addressed in the literature. The currently used VIM based on the error rate as a prediction accuracy measure does not seem suitable in the case of an ordinal response because the error rate does not differentiate between different kinds of misclassification. A classification of a healthy person as badly ill and a classification of a healthy person as slightly ill are regarded to be equally bad, though the latter is obviously a much better classification than the first. In the case of an ordinal response not all misclassifications can be regarded as equally poor and one might think about replacing the error rate by a more appropriate performance measure when computing the importance score of a predictor.

In this chapter, it is investigated whether incorporating the ordering information contained in the response improves RF's prediction performance and predictor ranking through RF. To improve predictor ranking for ordinal responses, the use of three alternative permutation VIMs is investigated, which are based on the mean squared error, the mean absolute error and the ranked probability score, respectively, that all take the ordering information into account. While the VIM based on the mean squared error is an established VIM that is frequently used for RF in the context of regression problems, the latter two VIMs are novel and have not been considered elsewhere. Finally, the impact of the choice of scores on prediction performance and on predictor rankings is explored. These issues are investigated using the RF version of Hothorn, Hornik and Zeileis (2006) as it is suitable for various kinds of regression problems, including ordinal regression.

3.2. Methods

3.2.1. Performance measures

In the following definitions of established performance measures are given. The performance measures are used in the studies for two purposes: i) to evaluate the prediction accuracy of RF for predicting an ordinal response and ii) for use in the proposed alternative permutation VIMs.

Error rate (ER) The error rate for the classification of observations $i = 1, \ldots, n$ with true classes Y_i into predicted classes $\hat{Y}_i$ is given by

$$ER = \frac{1}{n} \sum_{i=1}^{n} I(\hat{Y}_i \neq Y_i), \tag{3.1}$$

where $I(\cdot)$ denotes the indicator function. The error rate does not take the ordering of the classes into account since it only distinguishes between a correct classification ($\hat{Y} = Y$) and an incorrect classification ($\hat{Y} \neq Y$).

Mean squared error (MSE) With the mean squared error not all misclassifications are regarded as equally bad as is the case for the error rate. A higher penalty is put on a classification into a class which is more distant from the true class Y than on a classification into a class which is closer to Y. Let Y be an ordinal response that falls into ordered categories arbitrarily coded as $r = 1, \ldots, k$. To measure the distance between ordinal response classes scores $s(r) \in \mathbb{R}$ are used, with $s(1) < s(2) < \ldots < s(k)$. The distance between the true class Y and the predicted class $\hat{Y}$ is then computed from the difference in the corresponding scores, $s(\hat{Y}) - s(Y)$. Treating an ordinal variable as interval scaled by at-

tributing scores might be problematic. However, it has the advantage that loss functions for interval scaled variables like the mean squared error in the form

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(s(\hat{Y}_i) - s(Y_i))^2 \tag{3.2}$$

might be used (see e.g. Tutz (2011) p. 474, Fürnkranz and Hüllermeier (2010) p. 134, and Hechenbichler and Schliep (2004)). When using the simple scores $s(r) = r$, Eq. (3.2) yields $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$.

Mean absolute error (MAE) The mean absolute error used for the studies on ordinal regression problems is very similar to the mean squared error, with the difference that classification into a distant class is not penalized as much. Using the same notation as before, the mean absolute error for ordinal regression problems is given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|s(\hat{Y}_i) - s(Y_i)|. \tag{3.3}$$

For metric response Y the mean absolute error takes the form $\frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|$ which directly results from Eq. (3.3) when using the simple scores $s(r) = r$.

Ranked probability score (RPS) The ranked probability score originally introduced by Epstein (1969), is a generalization of the Brier score to multiple categories. It can be computed as the sum of Brier scores for all two-class problems that arise when splitting the sample on all possible thresholds made between two adjacent categories. The RPS has been shown to be particularly appropriate for the evaluation of probability forecasts of ordinal variables (Murphy; 1970). It is defined as

$$RPS = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k}(\hat{\pi}_i(r) - I(Y_i \leq r))^2, \tag{3.4}$$

where k denotes the number of response classes and $\hat{\pi}_i(r)$ denotes the predicted probability of observation i belonging to classes $\{1, \ldots, r\}$. The RPS measures the discrepancy between the predicted cumulative distribution function and the true cumulative distribution function (Murphy; 1970). The predicted cumulative distribution function can be computed from class probabilities that are predicted by a model, that is the estimated probabilities of an observation belonging to classes $r = 1, \ldots, k$. The true cumulative distribution function simplifies to a step function with a step from 0 to 1 at the true value Y_i for observation i. A graphical illustration of the RPS is given in Figure 3.1 for an observation i with observed category $y_i = 6$. Figure 3.1 shows the true cumulative distribution

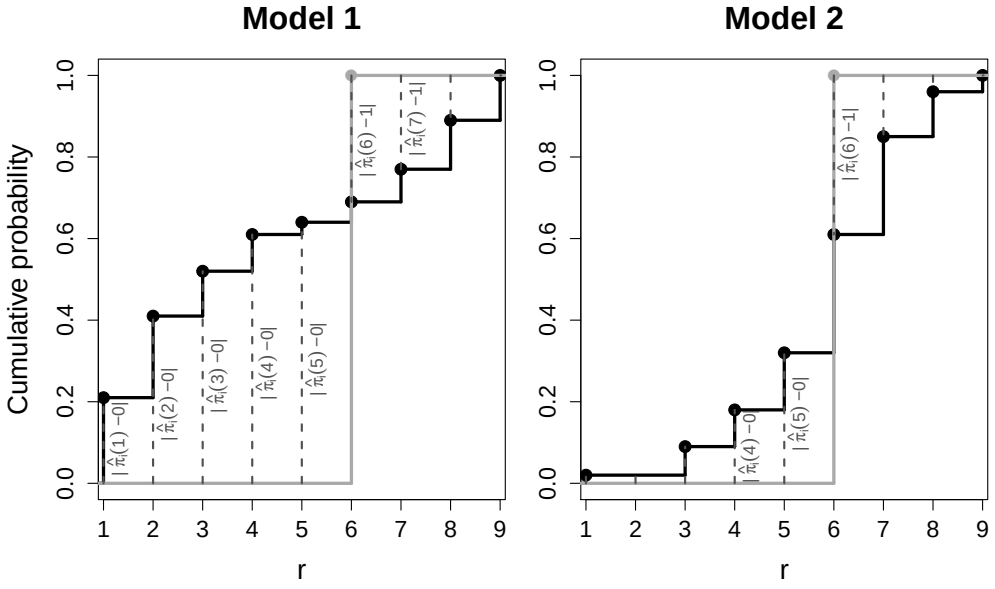

Figure 3.1.: Predicted (solid black line) and true (solid gray line) cumulative distribution functions for an individual with observed category $y_i = 6$ for two different models. Dashed lines indicate the difference between the predicted and the true cumulative distribution functions, that is $|\hat{\pi}_i(r) - I(y_i \leq r)|$, for $r = 1, \ldots, k$ and $y_i = 6$.

function (solid gray line) with step from 0 to 1 at the true value $y_i = 6$ and the cumulative distribution function (solid black line) that is obtained from class predictions of a model. Predicted distribution functions are given for two different models. The dashed lines correspond to the distance between the predicted and the true cumulative distribution functions (i.e., $\hat{\pi}_i(r) - I(6 \leq r)$) for a specific category r. These distances are squared when computing the RPS as in Eq. (3.4). The predicted cumulative distribution function in the left panel indicates that Model 1 does not seem to be very accurate in predicting the value for observation i. Here distances between the true and the predicted cumulative distribution functions are large and the RPS for observation i takes the value $0.21^2 + 0.41^2 + 0.52^2 + 0.61^2 + 0.64^2 + (0.69-1)^2 + (0.77-1)^2 + (0.89-1)^2 + (1-1)^2 = 1.4254$. A much better prediction is obtained when using Model 2. This model assigns the greatest probabilities for values of or around the true value $y_i = 6$. Accordingly, the distances between the true and the predicted cumulative distribution functions are rather small, which is reflected by an RPS of $0.02^2 + 0.02^2 + 0.09^2 + 0.18^2 + 0.32^2 + (0.61-1)^2 + (0.85-1)^2 + (0.96-1)^2 + (1-1)^2 = 0.3199$. It is clear from this illustration that the RPS is smaller (indicating a better prediction) if the predicted probabilities are concentrated near the observed class and is minimal if the predicted probability for the observed class is 1. From its definition it is clear that the RPS uses solely the ordering of the categories and does not require information on the distances between categories.

3.2.2. Novel variable importance measures for ordinal response

In the R package party (Hothorn et al.; 2012), the VIM for ordinal regression trees is the error rate based VIM. However, there are no studies that have shown that the error rate based VIM is suitable in the case of ordinal response and that it should be preferred over, for example, the MSE-based VIM. Two novel permutation VIMs are introduced which might be, in addition to the MSE-based VIM, promising for ordinal response data. These VIMs are based on the performance measures introduced in Section 3.2.1. More precisely, VIMs of the form (2.7) are proposed, where the ranked probability score (cf. Eq. (3.4)) or the mean absolute error (cf. Eq. (3.3)) are used as the error measure M. These VIMs will be termed the *RPS-based VIM* and the *MAE-based VIM*. While the error rate based VIM does not take the ordering information of the response levels into account, the three other VIMs do. In the studies the performances of the four VIMs are investigated and compared.

The implementation available at the website `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/janitza/rf_ordinal/index.html` allows the computation of the novel VIMs from either ordinal regression or classification trees, if constructed using the R package party. In addition to the RPS- and MAE-based VIMs, an implementation of the MSE-based VIM is provided that enables one to compute the MSE-based VIM from ordinal regression trees and from classification trees as well, a feature which is not currently possible using the R package party.

The computation of VIMs involves obtaining tree predictions for the OOB observations. The computational complexity for computing the predictions for a tree is directly related to the tree's depth. Louppe (2014) shows that in the worst case this is $\mathcal{O}(n^2)$, which corresponds to a tree where all splits put a single training observation in one daughter node, while all other training observations are put in the other daughter node. In the best case always half of the training observations are put in each daughter node, so that time complexity for obtaining tree predictions is at best $\mathcal{O}(n\log(n))$. According to Louppe (2014) "the analysis of the average case shows however that pathological cases are not dominant and that, on average, complexity behaves once again as in the best case." The computational complexity of the computation of the RPS is $\mathcal{O}(nk)$, so that for the average case the total time complexity for computing the RPS-based VIM amounts to $\mathcal{O}(n\log(n) + nk)$. Since the number of response classes k is usually much smaller than $\log(n)$, computing the RPS-based VIM is of order $n\log(n)$. The three other VIMs are of the same time complexity.

3.3. Simulation studies

3.3.1. Studies on prediction accuracy

Using the RF version based on conditional inference trees two RF variants were compared with respect to their ability to predict an ordinal response:

1. *RF ordinal*. RF consisting of ordinal regression trees. Simulations were performed using default scores (i.e., $s(r) = r, r = 1, \ldots, k$). Additional studies with quadratic scores $s(r) = r^2, r = 1, \ldots, k$, were also performed.

2. *RF classification*. RF consisting of classification trees. The ordinal response is treated as nominal, meaning that the information regarding the natural ordering of the levels of the response is ignored.

Prediction accuracy of a RF variant was assessed using the ranked probability score (RPS; see Eq. (3.4)) and the error rate (see Eq. (3.1)) computed for a large independent test data set of size $n = 10000$ that followed the same distribution as the training set on which the RFs were fit. Note that the RPS and the error rate do not necessarily come to the same conclusion, meaning that the error rate might be lower for one RF variant than for the other but its RPS is higher. Since the error rate does not consider how "severe" a misclassification is, the RPS is considered to be a more appropriate performance measure for evaluating a model that predicts an ordinal response. Thus the focus is on the results that are obtained when using the RPS for assessing prediction accuracy.

3.3.2. Studies on variable importance

Permutation VIMs based on the different performance measures described in Section 3.2.2 were applied to see which VIMs are most appropriate in the case of ordinal response. The four different VIMs were computed for RF constructed from ordinal regression trees (*RF ordinal*) as well as for RF using classification trees (*RF classification*; see Section 3.3.1).
VIMs give a ranking of the predictors according to their association with the response. To evaluate the quality of the rankings of the VIMs, the area under the curve (AUC) was used. Let the predictor variable indices $B = \{1, \ldots, p\}$ be partitioned into two disjoint sets $B = B_0 \cup B_1$, where B_0 represents the *noise predictors* (without any effect) and B_1 represents the *signal predictors* (with effect). The AUC is computed as follows:

$$AUC = \frac{1}{|B_0|\,|B_1|} \sum_{i \in B_0} \sum_{j \in B_1} I(VI_i < VI_j) + 0.5 I(VI_i = VI_j) \tag{3.5}$$

where $|B_l|$ denotes the cardinality of B_l with $l \in \{0, 1\}$, and $I(\cdot)$ denotes the indicator function (see, e.g., Pepe; 2004). Note that the AUC is often used for evaluating the abil-

ity of a method (which may be for example a diagnostic test or a prediction model) to correctly discriminate between observations with binary outcomes (often diseased versus healthy). In the studies, in contrast, the AUC is computed considering the predictor variables $X_1, \ldots, X_p$ as the units to be predicted (as noise or signal variables) rather than the observations $i = 1, \ldots, n$. The AUC here corresponds to an estimate of the probability that a randomly drawn signal predictor has a higher importance score than a randomly drawn noise predictor. Thus the AUC was computed in the studies to assess the ability of a VIM to differentiate between signal and noise predictors. An AUC value of 1 means that each of these signal predictors receives a higher importance score than any noise predictor, thus indicating perfect discrimination by the VIM. An AUC value of 0.5 means that a randomly drawn signal predictor receives a higher importance score than a randomly drawn noise predictor in only half of the cases, indicating no discriminative ability by the VIM.

3.3.3. Data simulation

The data were simulated from a mixture of two proportional odds models. Let $P(Y \leq r|\boldsymbol{x})$ denote the cumulative probability for the occurrence of a response category equal to or less than r for an individual with covariate vector $\boldsymbol{x}$. This probability is derived from a mixture of two proportional odds models

$$P(Y \leq r|\boldsymbol{x}) = \zeta P_1(Y \leq r|\boldsymbol{x}) + (1-\zeta)P_2(Y \leq r|\boldsymbol{x}), \tag{3.6}$$

where ζ is the mixture proportion and $P_1(Y \leq r|\boldsymbol{x})$ and $P_2(Y \leq r|\boldsymbol{x})$ are the cumulative probabilities that arise from two independent proportional odds models. The proportional odds model for mixture component $g \in \{1, 2\}$ has the form

$$P_g(Y \leq r|\boldsymbol{x}) = \frac{\exp(\gamma_{0rg} + \boldsymbol{x}^\top \boldsymbol{\gamma}_g)}{1 + \exp(\gamma_{0rg} + \boldsymbol{x}^\top \boldsymbol{\gamma}_g)}, r = 1, \ldots, k, \tag{3.7}$$

where the category-specific intercepts satisfy the condition $\gamma_{01g} \leq \ldots \leq \gamma_{0kg} = \infty$. In contrast to the intercepts, the coefficients $\boldsymbol{\gamma}_g$ do not vary over categories. In this case the comparison of two individuals with respect to their cumulative odds $P_g(Y \leq r|\boldsymbol{x})/P_g(Y > r|\boldsymbol{x})$ for mixture component g does not depend on the category r, giving the model its name, "proportional odds model" (see e.g., Tutz; 2011).

In the studies, the intercepts do not differ between the two mixture components; that is $\gamma_{0r1} = \gamma_{0r2} = \gamma_{0r}$. The intercepts for the categories were chosen such that the difference between the intercepts of adjacent categories is larger for more extreme categories. Concrete values for the intercepts are provided in Table 3.1.

Number of response levels	γ_{01}	γ_{02}	γ_{03}	γ_{04}	γ_{05}	γ_{06}	γ_{07}	γ_{08}	γ_{09}
$k = 3$	-1.80	1.80	∞	-	-	-	-	-	-
$k = 6$	-4.50	-1.50	0.00	1.50	4.50	∞	-	-	-
$k = 9$	-5.90	-3.41	-1.55	-0.31	0.31	1.55	3.41	5.90	∞

Table 3.1.: Intercepts for the proportional odds model (3.7) with $\gamma_{0rg} = \gamma_{0r}$.

The simulation setting comprises both predictors not associated with the response (noise predictors) and associated predictors (signal predictors). Predictors $X_1, \ldots, X_{15}$ had an effect on the cumulative odds of the first mixture component. The first five predictors each had a large effect, with corresponding parameter coefficients $\gamma_{1j} = 1$ for $j = 1, \ldots, 5$; the second set of five predictors each had a moderate effect, with coefficients $\gamma_{1j} = 0.75$ for $j = 6, \ldots, 10$; and the last set of five signal predictors each had a small effect, with coefficients $\gamma_{1j} = 0.5$ for $j = 11, \ldots, 15$. The remaining predictors $X_{16}, \ldots, X_{65}$ had no effect on the cumulative odds of the first mixture component and their respective coefficients were zero. For the second mixture component fewer predictors had an effect but all effects were large (coefficient of either 1 or -1). Almost all predictors which had an effect for the first component, had an effect for the second – with the exceptions of X_5, X_{10} and X_{15}, which had no effect for the second component. For predictors $X_5, X_{10}, X_{15}, X_{16}, X_{17}, \ldots, X_{65}$ the corresponding coefficients were set to zero, while for the other predictors the parameter coefficients were $\gamma_{2j} = 1$ for $j \in \{1, 2, 6, 7, 11, 12\}$ and $\gamma_{2j} = -1$ for $j \in \{3, 4, 8, 9, 13, 14\}$. Table 3.2 shows the coefficients for both mixture components. To summarize, there are predictors that have no effect at all, predictors that have an effect for both mixture components and predictors that have an effect for only one mixture component.

Mixture component	Coefficient vector $\gamma_g^\top = (\gamma_{g1}, \ldots, \gamma_{g,65})$
$g = 1$	$(1, 1, \ 1, \ \ 1, 1, 0.75, 0.75, 0.75, 0.75, 0.75, 0.5, 0.5, 0.5, 0.5, 0.5, 0, 0, \ldots, 0)$
$g = 2$	$(1, 1, -1, -1, 0, 1, \ \ 1, \ \ -1, \ -1, \ 0, \ \ 1, \ 1, \ -1, \ -1, 0, \ 0, 0, \ldots, 0)$

Table 3.2.: Effects of predictors on the cumulative odds of the proportional odds model (3.7) for mixture components $g = 1, 2$.

Data was generated for sample sizes $n = 200$ and $n = 400$. Let $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{i,65})$ denote the covariate vector for the observation i. For the generation of the response value y_i the cumulative probability for the occurrence of a response category equal to or less than r was computed according to (3.6). Probabilities for classes $r = 1, \ldots, k$ were derived and a multinomial experiment was performed for each observation using its response class probabilities.

For each setting (specified in the following) 100 data sets were generated.

A further study was conducted to see if results differ in high-dimensional data settings. This study is shown in Appendix C.1.

Simulation settings

Various settings were simulated that differed in

- the value for the mixture proportion ζ. Settings were simulated for $\zeta = 0.6$ (data generation based on a mixture of two proportional odds models), $\zeta = 1$ (data generation based on the proportional odds model specified by mixture component $g = 1$) and $\zeta = 0$ (data generation based on the proportional odds model specified by mixture component $g = 2$),
- the number of ordered response levels, chosen as $k = 3$, $k = 6$ and $k = 9$, and,
- the generation of predictor variables. For settings without correlations, $\boldsymbol{x}_i, i = 1, \ldots, n$, were drawn from $\mathrm{N}(\mathbf{0}_p, \boldsymbol{I}_p)$, with $\boldsymbol{I}_p$ denoting the identity matrix of dimension $p \times p$ and p denoting the number of predictors. For settings with correlations, $\boldsymbol{x}_i, i = 1, \ldots, n$, were drawn from $\mathrm{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_p)$ with block diagonal covariance matrix

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} \boldsymbol{A}_{\text{signal}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{A}_{\text{noise}_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{A}_{\text{noise}_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_4} & 0 \\ 0 & 0 & 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_5} \end{bmatrix}.$$

The first block matrix $\boldsymbol{A}_{\text{signal}} \in \mathbb{R}^{(15\times 15)}$ determined the correlations among the signal predictors $X_1, \ldots, X_{15}$. It was defined as $\boldsymbol{A}_{\text{signal}} = (a_{ij})$ with

$$a_{ij} = \begin{cases} 1, & i = j \\ 0.8, & i \neq j; i, j \in \{1, 3, 6, 8, 11, 13\} \\ 0, & \text{otherwise} \end{cases}$$

in this way generating uncorrelated and also strongly correlated signal predictors. The matrices $\boldsymbol{A}_{\text{noise}_j} \in \mathbb{R}^{(10\times 10)}$ for $j = 1, \ldots, 5$ were given by

$$\boldsymbol{A}_{\text{noise}_j} = \begin{bmatrix} 1 & \rho_j & \ldots & \rho_j \\ \rho_j & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_j \\ \rho_j & \ldots & \rho_j & 1 \end{bmatrix},$$

and determined correlations among a set of 10 noise predictor variables with $\rho_1 = 0.8$, $\rho_2 = 0.6$, $\rho_3 = 0.4$, $\rho_4 = 0.2$ and $\rho_5 = 0$.

Random forests parameter setting

Simulation studies were performed using the unbiased RF version based on conditional inference trees which is implemented in the R package party (Hothorn et al.; 2012). For the studies, the setting for unbiased tree construction was used as suggested by Strobl et al. (2007). In this setting subsamples of size $0.632n$ are used instead of bootstrap samples in order to avoid possible biases induced by the bootstrap. Further, no p-value threshold is applied when implementing a split (i.e., the significance threshold α is set to 1 in step 1 of the algorithm; see p. 13). No other stopping criteria such as a minimum number of observations in a terminal node or a minimum number of observations required for a node to be split were applied. The number of randomly drawn predictors *mtry* was set to the default value $\lfloor \sqrt{p} \rfloor$, where p denotes the total number of predictors (here $p = 65$). The number of trees *ntree* was set to 1000.

3.3.4. Results

In the following the results of the simulation studies for the sample size of $n = 200$ are shown. The results for $n = 400$ are similar and thus not shown.

Prediction accuracy

Figure 3.2 shows the results of the simulation studies on the comparison of *RF ordinal* and *RF classification* with respect to their predictive accuracy (measured in terms of RPS). For a direct comparison, the ratio of the RPS for *RF ordinal* to that for *RF classification* is shown. Values of the RPS ratio below 1 mean that the prediction error as measured by RPS is smaller for *RF ordinal* and thus are in favor of *RF ordinal*. Conversely, values above 1 mean that the prediction error as measured by RPS is larger for *RF ordinal* and advocate the use of *RF classification* for prediction purposes. For values close to 1 prediction accuracies of *RF ordinal* and *RF classification* are comparable. In all settings the ratio of RPS is in the range $[0.92; 1.04]$ and thus is very close to 1, so there are no large differences between the prediction accuracies of the forest types in the simulation studies. However, one can observe a trend toward better performance of *RF ordinal* for a larger number of response levels. Overall, the performance is better for *RF ordinal* in most of the settings, except for $k = 3$, in which the performance of *RF classification* is better in two of six settings. Similar results were obtained when performance was measured in terms of the error rate. Further, the results generalize to high-dimensional data settings, as shown by simulations

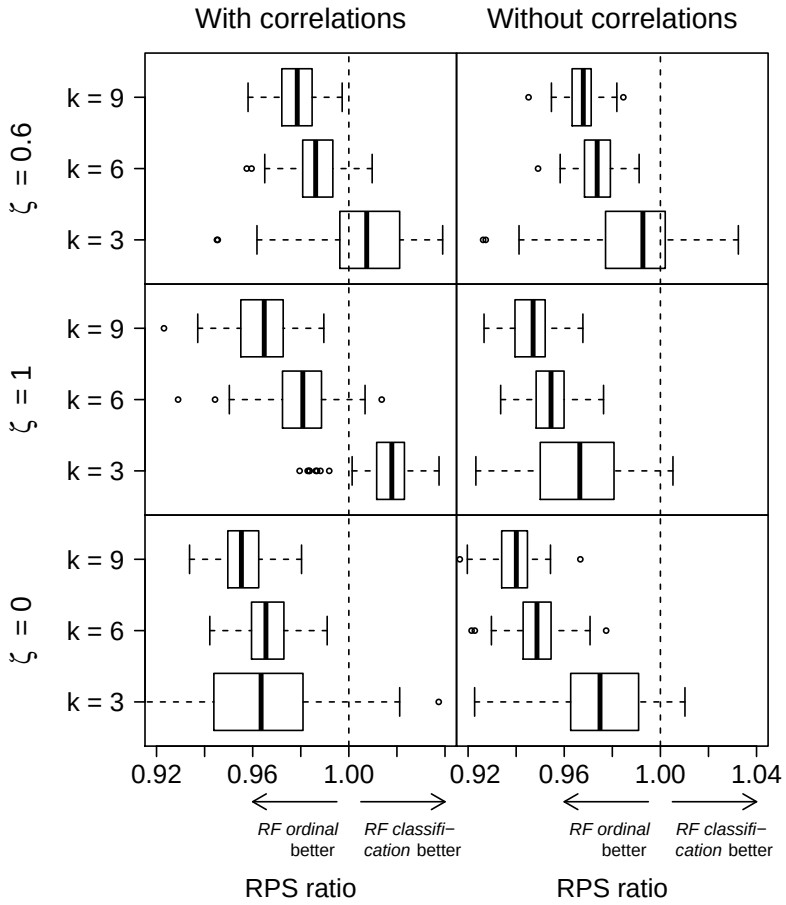

Figure 3.2.: Performance ratio for *RF ordinal* versus *RF classification* for simulated data. A ratio of the ranked probability scores (RPS) below 1 indicates a better prediction accuracy of *RF ordinal* and a ratio above 1 indicates a better prediction accuracy of *RF classification*. Data was generated for $n = 200$ from a mixture of proportional odds models (3.6) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row). Data was generated for $k \in \{3, 6, 9\}$ ordered response levels and for settings in which predictors correlate (left column) and in which all predictors are uncorrelated (right column).

in which the number of candidate predictors is larger than the number of observations (see Appendix C.1). Note that the results presented here were obtained by using equally spaced scores. The results are very similar when using quadratic scores, which suggests that the conclusions do not depend on the specific choice of scores for *RF ordinal*.

Performance of variable importance measures

Figures 3.3 - 3.5 show the results of the simulation studies on the performance of VIMs when using the four VIMs outlined in Sections 3.2.2 and 2.2.2, computed for both *RF ordinal* and *RF classification*. Here only the results are shown when using default (i.e., equally spaced) scores for tree construction and MSE- and MAE-based VIM computation. Very similar results were obtained when specifying quadratic scores.

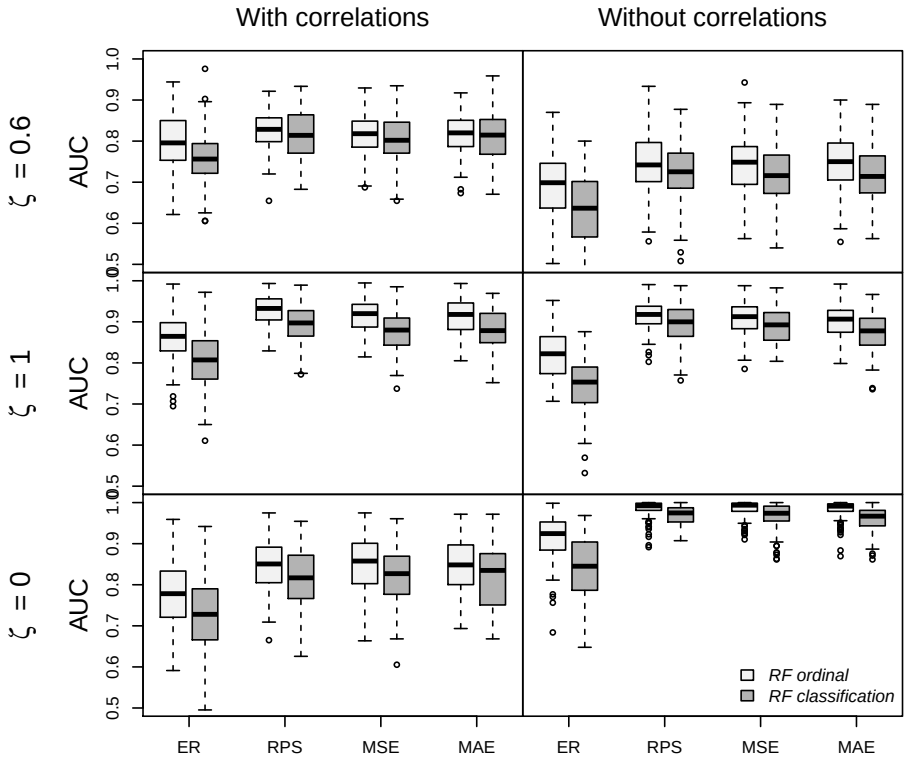

Figure 3.3.: Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 9-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models (3.6) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

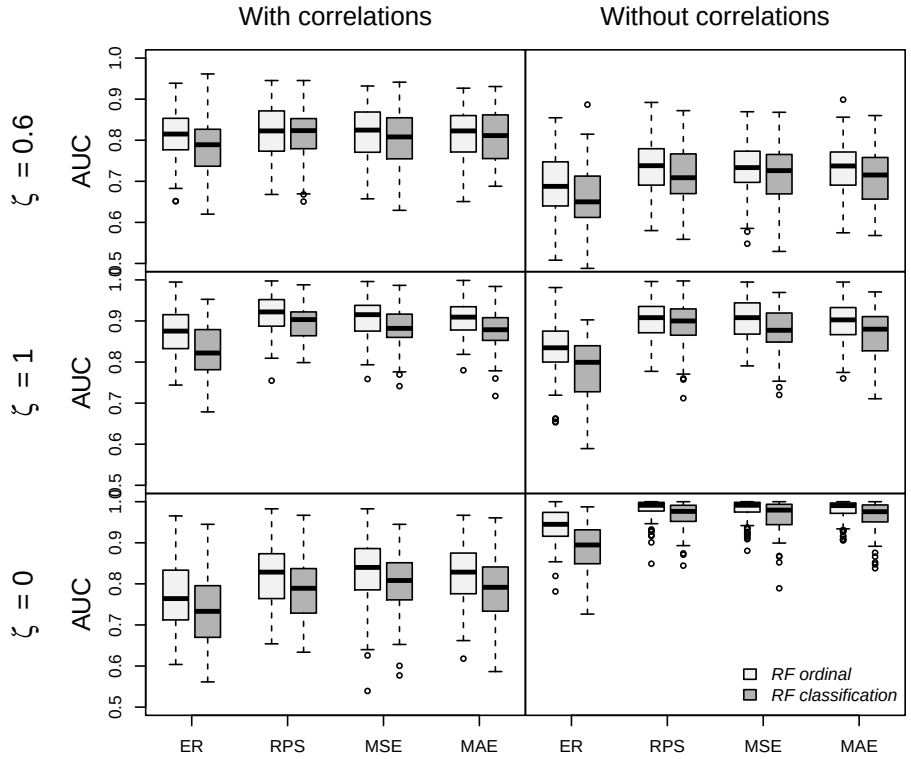

Figure 3.4.: Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 6-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models (3.6) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

In the settings with 9 response levels (Figure 3.3) the performances of the MSE-based VIM and the two novel permutation VIMs are consistently better than that of the error rate based VIM, independent of the type of trees used (ordinal regression or classification trees). Obviously, making use of the ordering is advantageous when deriving the importance of variables for these settings. Interestingly, in some settings the difference is rather small and in others it is more pronounced. Similar results are obtained for the setting with 6 response levels (Figure 3.4). However, the difference between the error rate based VIM and the other VIMs is less pronounced than for the settings with a 9-category response variable. In all settings in which the response has only 3 levels, the differences between the VIMs are marginal (Figure 3.5), though overall the novel VIMs and the MSE-based VIM remain superior. In the studies the RPS-based, MSE-based and MAE-based VIM show comparable performances.

The results suggest that the performances of all VIMs can in some settings be further improved by making use of the ordering in the construction of trees, through the application of ordinal regression trees. If used in combination with ordinal regression trees, the novel VIMs and the MSE-based VIM achieved the most accurate predictor rankings. The

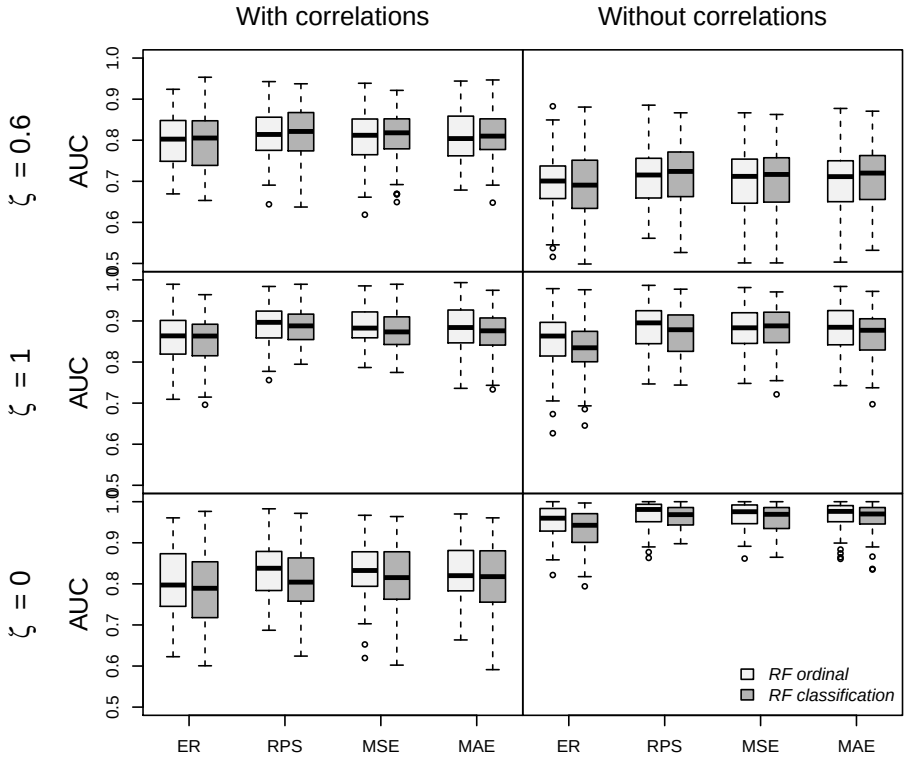

Figure 3.5.: Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 3-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models (3.6) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

worst rankings in contrast were obtained for the classical error rate based permutation VIM (which is currently in use for ordinal responses in the R package party) computed from classification trees. This indicates that predictor rankings are worst when making no use of the ordering at all, neither in tree construction nor in the computation of the variables' importance scores. Similar results were obtained for the high-dimensional data setting shown in Appendix C.1.

A plausible explanation for the improvement in the ranking by using ordinal regression trees is that in ordinal regression trees it is more likely that a predictor associated with the response is selected for a split. A predictor that is often selected in a tree and occurs close to the root node of the tree is likely to receive a high importance score. The advantage when applying ordinal regression trees is that the power of the statistical test to correctly detect an association between a predictor and the ordinal response is higher. It is thus less likely that a noise predictor yields a lower p-value just by chance and is selected for the split. Results obtained for the described simulation studies provide evidence for this. One can, for example, inspect the trees of a forest and compute the number of trees for which an influential predictor was chosen for the first split. If the fraction of

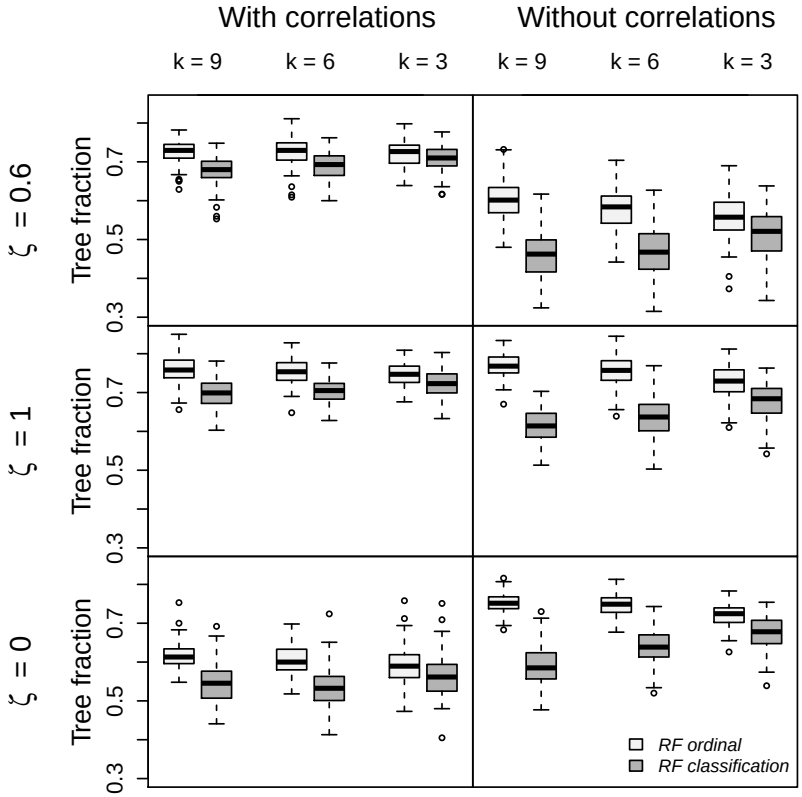

Figure 3.6.: Fraction of trees in *RF ordinal* and *RF classification* where an influential predictor was selected for the first split. Distributions arise from 500 replications of the simulation setting described in Section 3.3.1 with $k = 3$ response levels, $k = 6$ and $k = 9$. Data was generated for $n = 200$ using a mixture of proportional odds models (3.6) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

trees is significantly higher for the forest consisting of ordinal regression trees, this is an indication that ordinal regression trees are more accurate in selecting predictors for a split compared to classification trees. For the simulation studies the fraction of trees where a signal predictor was selected for the first split was calculated for both *RF ordinal* and *RF classification*; the results are displayed in Figure 3.6. The results confirm the hypothesis that *RF ordinal* is more accurate in selecting important predictors for a split than *RF classification*. Since the power of a test that takes into account the ordering increases with the number of ordered categories, the discrepancy between *RF ordinal* and *RF classification* is most pronounced for $k = 9$ and least pronounced for $k = 3$.

3.4. Real data applications

In the studies five publicly available real data sets with an ordinal response were considered. Note that the results for all data sets that were analyzed are reported, so there was no selection of the data sets depending on the obtained results.

3.4.1. Data

The Very Low Birth Weight data was analyzed by O'Shea et al. (1998) for identifying perinatal events from sonographical and echodensity measurements. The data can be obtained from the website `http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets`. In the analyses the Apgar score (a score for the physical health status of a newborn measured on a 9-point scale) was supposed to be predicted from diverse factors such as medication the mother took during pregnancy, weight and sex of the newborn and the type of delivery.

The Wine Quality data is available from the UCI repository (`http://archive.ics.uci.edu/ml/datasets.html`); see also Cortez et al. (2009) for details on the data. The response to be predicted from physicochemical measurements (like alcohol concentration or residual sugar) was the quality of a wine, measured on a scale from 0 (poorest quality) to 10 (highest quality). There were no observations with the highest quality (i.e., a score of 10) and very poor quality (score from 0 - 2). Due to their small number ($n = 5$), observations with a score of 9 were removed from the data.

The National Health and Nutrition Examination Survey (NHANES) is a series of cross-sectional surveys of the US population (National Center for Health Statistics; 2012). The data can be obtained from the institution's homepage. An overview of the considered data is given in Appendix B. The self-reported general health status was considered as the outcome variable to be predicted from demographical and health-related factors. The response is categorized into five categories (1: excellent, 2: very good, 3: good, 4: fair, 5: poor).

The data for the SUPPORT Study can be obtained from the website `http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets`. The considered data set is a random sample of 1000 patients from phases I & II of the **S**tudy to **U**nderstand **P**rognoses and **P**references for **O**utcomes and **R**isks of **T**reatment (SUPPORT) (Knaus et al.; 1995). Several outcomes in seriously ill hospitalized adults have been considered. A focus was on the prediction of functional disability, which is categorized into 5 ordered categories from slight to severe (see Table 3.3 for details).

The Mammography Experience data was analyzed by Hosmer Jr and Lemeshow (2004) (p. 264), who studied the relationship between mammography experience (have never had a mammography, have had one within the last year, last mammography greater than one year ago) and the attitude toward mammography based on a study questionnaire. The data is part of the R package TH.data.

For all data sets (except for the Very Low Birth Weight data) covariates for which more than 10% of the observations had missing values, were excluded. Observations with missing values in any of the included covariates were deleted. An overview of the num-

Data	Considered response variable	Levels
Very Low Birth Weight	Apgar score	1 (life-threatening) ($n = 33$) 2 ($n = 16$) 3 ($n = 19$) 4 ($n = 15$) 5 ($n = 25$) 6 ($n = 27$) 7 ($n = 35$) 8 ($n = 36$) 9 (optimal physical condition) ($n = 12$)
Wine Quality	Wine quality score[#]	3 (moderate quality) ($n = 20$) 4 ($n = 163$) 5 ($n = 1457$) 6 ($n = 2198$) 7 ($n = 880$) 8 (high quality) ($n = 175$)
NHANES	Self-reported health status	1 – excellent ($n = 198$) 2 – very good ($n = 565$) 3 – good ($n = 722$) 4 – fair ($n = 346$) 5 – poor ($n = 83$)
SUPPORT Study	Functional disability	1 – patient lived 2 months, and from an interview (taking place 2 months after study entry) there were no signs of moderate to severe functional disability ($n = 310$) 2 – patient was unable to do 4 or more activities of daily living 2 months after study entry; if the patient was not interviewed but the patient's surrogate was, the cutoff for disability was 5 or more activities ($n = 104$) 3 – Sickness Impact Profile total score is at least 30 2 months after study entry ($n = 57$) 4 – patient intubated or in coma 2 months after study entry ($n = 7$) 5 – patient died before 2 months after study entry ($n = 320$)
Mammography Experience	Last mammography visits	1 – never ($n = 234$) 2 – within a year ($n = 104$) 3 – over a year ($n = 74$)

Table 3.3.: Response variables of the five real data sets and their frequency in the analyzed data.
[#] There were no observations with categories 0, 1, 2, 9, 10 in the analyzed data set.

Data	No. response levels k	No. predictors p	No. observations n
Very Low Birth Weight	9	10	218
Wine Quality	6	11	1599
NHANES	5	26	1914
SUPPORT Study	5	16	798
Mammography Experience	3	5	412

Table 3.4.: Number of response levels, predictor variables and observations for the considered data sets.

ber of response levels, predictor variables and observations for the data sets (as used for the analysis) is given in Table 3.4. Table 3.3 gives an overview of the response variables. Note that there were types of responses ranging from different scoring systems (Wine Quality data, NHANES data and Very Low Birth Weight data), to categorizations of functional disability (SUPPORT Study), to the recentness of events, as grouped into 3 categories (Mammography Experience data).

All RF parameters were defined as described for the simulated data in Section 3.3 ($mtry = \lfloor \sqrt{p} \rfloor$, $ntree = 1000$, no early stopping). Default (i.e., equally spaced) scores were used in the analysis.

3.4.2. Studies on prediction accuracy

The ranked probability score (RPS; see Eq. (3.4)) and the error rate (see Eq. (3.1)) were used to assess prediction accuracies by *RF ordinal* and *RF classification*. Prediction accuracies were assessed using 10-fold cross-validation. The cross-validation was repeated 500 times to obtain more stable results.

3.4.3. Studies on variable importance

When using real data one usually faces the problem that it is unknown which of the variables are important and which are not. As is known from further investigations (not shown), for all five data sets there are at least some variables which improve response prediction since the predictions by the constructed forests were always more accurate than the predictions by the null model (i.e., that without covariates). If we assume that there was an additional set of variables which are not associated with the response, one would be able to investigate and compare the discriminative abilities of the VIMs: a well-performing VIM is namely expected to attribute higher importance scores to the original (and potentially important) predictors than to the noise predictors.

The following steps were implemented to study the performance of VIMs:

- The original data was augmented by a set of noise predictors. This was done by

duplicating the set of original predictor variables and then randomly permuting the rows of this duplicated predictor set. This ensured that each predictor within this duplicated predictor set was unrelated to the response variable, while preserving realistic correlation structures within the duplicated predictor set.

- *RF ordinal* and *RF classification* were fit to this augmented data and the variables' importance scores were derived using each of the four permutation VIMs described in Sections 2.2.2 and 3.2.2.

- The area under the curve (AUC) was used to measure the performance of VIMs. The AUC is an estimate of the probability that a randomly drawn predictor from the original (i.e., unpermuted) set of predictors would obtain a higher importance score than a randomly drawn predictor from the permuted set of predictors (see Section 3.3.2).

This process was repeated 500 times. Note that while in Section 3.3.2 an AUC value of 1 indicated perfect discrimination between signal and noise predictors, here it is expected that perfect discrimination can already be obtained for AUC values lower than 1: since it is likely that not all of the original variables are truly influential predictors, some of them actually should be regarded as noise predictors instead. However, this does not pose a problem for the studies because the aim is to *compare* the VIMs with respect to discriminative ability, so their absolute AUC values are not of interest but only the differences in the AUC values.

3.4.4. Results

Prediction accuracy

The results on prediction accuracy of *RF ordinal* and *RF classification* based on the five real data sets are shown in Figure 3.7. For a direct comparison of *RF ordinal* and *RF classification* the RPS ratio (left panel) and the error rate ratio (right panel) were computed. The results shown in Figure 3.7 are in line with the results obtained from the simulation studies in Section 3.3.4; overall the differences in prediction accuracies are rather small. The ratios are even closer to 1 than the ratios obtained for the simulated data (cf. Figure 3.2). In contrast to the simulated data, there is no trend with respect to the number of response levels. Instead, which RF variant performs better seems to depend highly on the considered data set as well as on which performance measure is used; when using the RPS as the performance measure (which is considered to be more appropriate than the error rate) for three of the data sets (Wine Quality data, NHANES data, Mammography Experience data) an at least marginally better predictive accuracy was obtained by *RF ordinal*, while

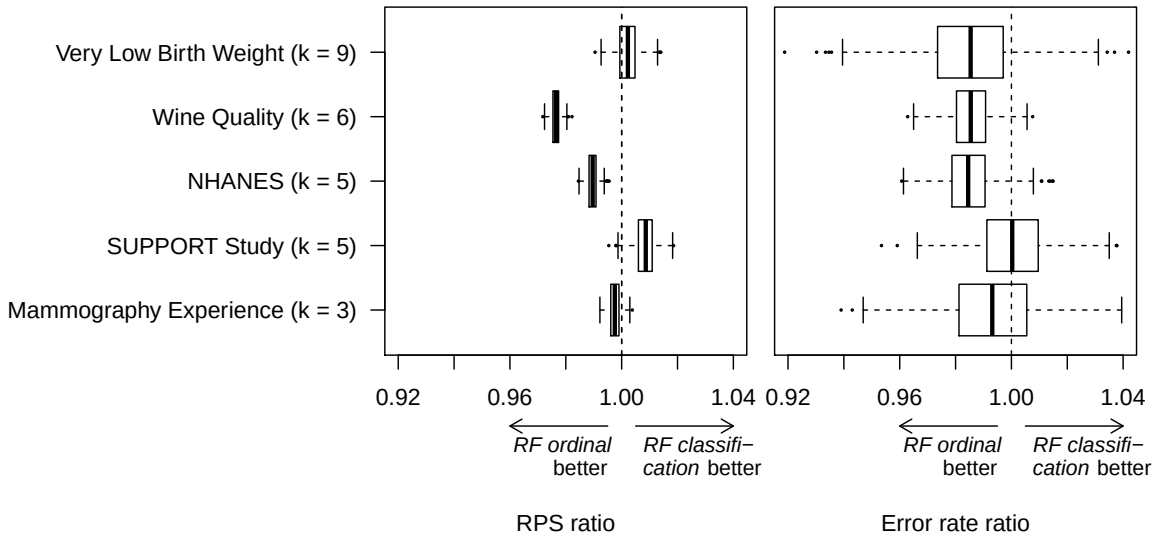

Figure 3.7.: Performance ratio for *RF ordinal* versus *RF classification* for the five real data sets. Values below 1 indicate a better performance of *RF ordinal* and values above 1 indicate a better performance of *RF classification*. Prediction accuracy was measured by ranked probability score (left) and error rate (right) using 10-fold cross-validation repeated for 500 random splits.

for the other two data sets (the Very Low Birth Weight data and the SUPPORT Study) *RF classification* gave slightly more accurate predictions. In contrast, *RF ordinal* is for all data sets at least as good as *RF classification* when the error rate is used as the performance measure.

Performance of variable importance measures

Figure 3.8 shows the AUC values over the 500 repetitions. Very marginal differences in performance can be observed when the importance scores of variables are derived from ordinal regression trees compared to classification trees. The performance of a VIM seems to depend highly on the nature of the response variable since results differ between the data sets. While for the Very Low Birth Weight data and for the NHANES data all three VIMs that take into account the ordering in response levels have better discriminative ability than the error rate based VIM, there is hardly any difference between the error rate based VIM and the two novel VIMs (based on the RPS and MAE) for the other three data sets. Note that for the Wine Quality data perfect discrimination for all VIMs is obtained, which indicates that all variables in the original data set are associated with the quality of a wine. Interestingly, in these studies, compared to the two novel VIMs based on the RPS and the MAE, the MSE-based VIM always performs worse or has equal performance at best.

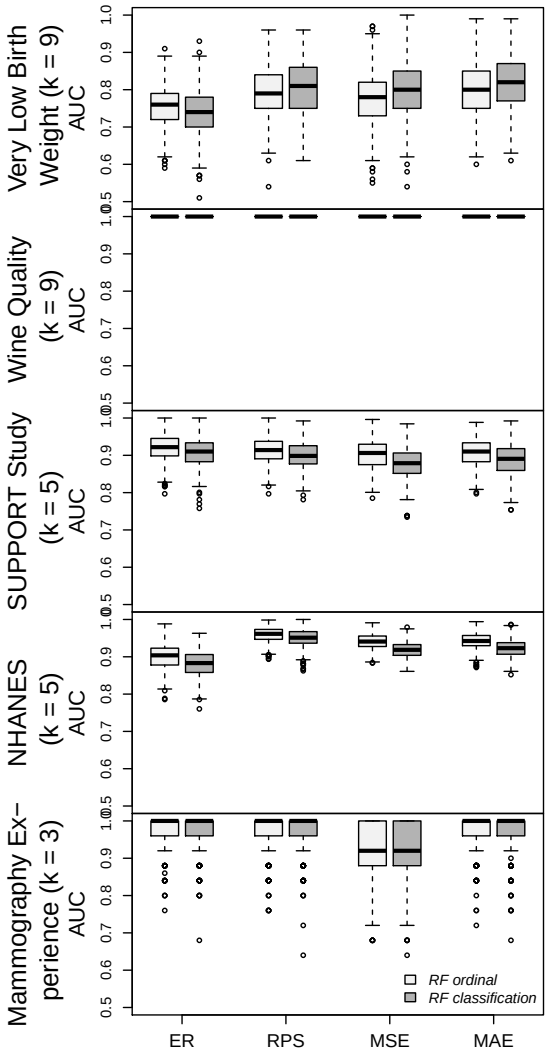

Figure 3.8.: Performance of different VIMs for five real data sets when computed on *RF ordinal* and *RF classification*. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). The performance of VIMs is measured in terms of the area under the curve (AUC), which corresponds to the probability that a randomly drawn potentially important predictor has a higher importance value than a randomly drawn noise predictor.

3.5. Discussion

The use of the ordering of the levels of an ordinal response variable in tree construction is not supported by the classical RF version of Breiman (2001). In practice, data with ordinal responses have often been handled using classification or regression trees. However, the former fully ignores the ordering and the latter assumes the response to be measured on a metric scale and yields metric values instead of class predictions. The RF implementation of Hothorn, Hornik and Zeileis (2006) in contrast, allows the modeling of various kinds of regression problems, including nominal, ordinal, numeric, and censored, as well as multivariate response variables and arbitrary measurement scales of the covariates. It is thus promising for applications in which the response has an inherent ordering. Moreover, this version is based on a conditional inference framework and, in contrast to the classical RF version of Breiman (2001), implements unbiased split selection. For these reasons the studies are based on the RF version of Hothorn, Hornik and Zeileis (2006).

In this chapter, it was investigated whether prediction accuracy improves when making use of the ordering of the levels of the response variable. For this purpose, using simulated and real data, the prediction accuracy of RF composed of classification trees was compared to that of RF composed of ordinal regression trees (i.e., trees for ordinal responses as implemented in the R package party; Hothorn, Hornik and Zeileis; 2006). The studies indicate that there are only small differences in prediction accuracy. For 16 of 18 studies based on simulated data and for 3 of 5 studies based on real data, more accurate class predictions were obtained for RF consisting of ordinal regression trees, suggesting that ordinal regression trees are a reasonable alternative to classification trees if the response is ordinal. However, the differences were only small and their practical relevance is questionable.

Prediction accuracy was primarily assessed by using the ranked probability score. The results were also investigated when prediction accuracy was evaluated based on the error rate and two alternative measures which are described in Appendix C.1. The results obtained with the error rate and the two alternative measures were consistent with the findings. Thus the conclusions do not depend on the choice of the accuracy measure.

In addition to prediction accuracy, it was also investigated if making use of the ordering for the computation of VIMs leads to more accurate predictor rankings. In the presence of an ordinal response the current RF implementation of Hothorn, Hornik and Zeileis (2006) uses the error rate based permutation VIM. Two novel permutation VIMs for RF that are promising in settings in which the response has an inherent ordering were introduced. The results on simulated and on real data showed that a VIM which makes use of the ordering in the levels of the response yields in many cases an at least slightly more accurate predictor ranking than the classical error rate based VIM, and thus should

be used when analyzing ordinal response data. The studies suggest that by using ordinal regression trees a further improvement in the predictor rankings might be obtained. This is most likely related to the fact that ordinal regression trees more often select relevant predictors for a split than classification trees since hypothesis tests used for split selection in conditional inference trees have higher statistical power for the detection of relevant effects if making use of the ordering of the response levels. In data settings where the response variable is ordinal it is thus recommended using a permutation VIM which makes use of the ordering in combination with ordinal regression trees if the aim is to obtain a predictor ranking or to select important variables. Among the VIMs that make use of the ordering, the two novel VIMs outperformed the well-known MSE-based VIM on real data.

Though in the studies on the performance of VIMs the RF version of Hothorn, Hornik and Zeileis (2006) was exclusively used, it is expected that VIMs that make use of the ordering, like the RPS-based VIM, give more accurate rankings also when using the classical RF version of Breiman (2001).

Note that the MSE-based VIM was developed for regression trees but had not been considered for ordinal responses to this point. While the RPS-based VIM relies only on the ordering of the levels, the MAE- and MSE-based VIMs require the specification of distances between the response levels. The two specific choices of the scores (reflecting distances in response levels) did not impact the results on the performance of variable importance measures or on prediction accuracies of the ordinal regression trees. This suggests that specific values for the scores do not seem to have a significant impact as long as the scores reflect the correct ordering of the levels. Though in the simulation studies different scores did not lead to different results, one cannot be sure that this also applies to other settings. Thus it is recommended to use the RPS-based VIM – which does not make any assumptions on the distance between response levels – over the MAE- and MSE-based VIMs.

Note that the incorporation of the ordering of the response levels was investigated when constructing trees and when computing the importance of variables. The ordering of the response levels in the context of another stage could also be considered in future studies, namely when aggregating tree predictions to obtain a final prediction of a class (see, e.g., Tutz; 2011, Section 15.9); in the context of k-nearest-neighbors it has for example been shown that such a procedure might give more accurate predictions (Hechenbichler and Schliep; 2004).

4. A variable importance test for high-dimensional data

This chapter presents a new heuristic method for testing RF's permutation VIM in high-dimensional data settings. This method was introduced in the paper of Janitza et al. (2015), for which I received the "Student Postdoctoral Fellow Paper Competition and Travel Award" by the IFCS. The current chapter is based on Janitza et al. (2015) and is structured as follows: After an introduction to the use of VIM tests in the medical literature, the heuristic testing approach of Altmann et al. (2010) is briefly reviewed in Section 4.2. The new heuristic testing idea is subsequently introduced. As will be shown, the testing idea is based on presumptions which are not met by the classical permutation VIM. Therefore a modified version of the permutation VIM is developed which fulfills the criteria and might be used in the testing procedure. Finally, the design of simulation studies is described, in which the novel testing approach is compared to the approach of Altmann et al. (2010) and to a naive approach which consists in applying the testing idea to the classical permutation VIM. Section 4.3 shows the results of the simulation studies and Section 4.4 gives a brief summary and discussion of the results.

4.1. Introduction

Often, identifying relevant genes is of high interest to gain valuable insights into the functionality and mechanisms that lead to a specific disorder. Moreover, the identification of relevant genes aids in the diagnosis of certain disorders. The RF method and its implemented VIMs have often been used for the identification of such biomarkers (e.g., Reif et al.; 2009; Wang-Sattler et al.; 2012; Yatsunenko et al.; 2012). There are two commonly used VIMs, the Gini VIM and the permutation VIM. While the Gini VIM has undesirable properties, the permutation VIM is essentially unbiased (see Section 2.2.2). The permutation VIM reflects the average decrease in accuracy when destroying the association between a variable and the response by permuting the values of the variable. It is clear that predictor variables whose importance score is negative or zero are likely to have no predictive ability. However, for the predictor variables with positive importance score

it is difficult to say which importance scores are large enough so that it is unlikely that these have occurred by chance. The VIM depends on several different factors, including factors related to the data, such as correlations between the data, the signal-to-noise ratio or the total number of variables, and including RF specific factors, such as the choice of the number of randomly drawn candidate predictor variables for each split. Therefore there is no universally applicable threshold that can be used to determine what really high importance scores are.

Often, in practical applications, a certain percentage of the highest ranked variables are selected; Reif et al. (2009) for example filtered out the 10% of variables with the highest importance scores and used them for further considerations. However, one should be careful when selecting a prespecified number of highest ranked variables and considering these as relevant because one would always identify some variables as relevant even in the absence of any associations between the variables and the response.

An ad-hoc approach consists in using the absolute value of the smallest observed negative importance score as a threshold for determining which variables are likely to be relevant, because one can be sure that the smallest observed negative importance score must have been occurred due purely to chance (Strobl et al.; 2009). However, this approach has several disadvantages, two of them being that the threshold depends on one single observed importance score and that it becomes more extreme the more variables there are. It is thus clear that more elaborated approaches are needed.

Testing procedures are a sensible strategy for deciding which variables are likely to be relevant (Huynh-Thu et al.; 2012). In a statistical test one aims to draw conclusions about the value of a population parameter through the use of the observed sample. In the context of VIMs it is not clear what this population parameter refers to and if it even exists. Thus the testing approaches that were proposed for RF's VIMs, should rather be regarded as heuristic methods that enable the selection of variables, instead of real statistical tests in the strict mathematical sense. However, for simplicity and to be consistent with the literature, such approaches will be referred to as statistical tests, although it should be kept in mind that in the strict mathematical sense these are not statistical tests.

A statistical test based on the supposed normality of a scaled version of the permutation VIM was proposed by Breiman and Cutler (2004). However, the procedure of Breiman and Cutler (2004) has been shown to have alarming statistical properties, and should not be used (Strobl and Zeileis; 2008). During the last years, more and more approaches have been developed that test which variables are related to the response (see Hapfelmeier and Ulm; 2013, and references therein). Since the true null distribution of variable importance depends on various factors, it becomes difficult – if not impossible – to theoretically derive the null distribution. This is the reason for the frequent use

of permutation strategies in the existing testing approaches (Tang et al.; 2009; Altmann et al.; 2010; Hapfelmeier and Ulm; 2013). However, such procedures are computationally demanding. Very recently Hapfelmeier and Ulm (2013) published a comprehensive comparison study of different permutation-based testing approaches. They conclude that their novel approach has higher statistical power than many of the existing approaches and controls the type I error. Their approach works as follows: For each variable that is tested for its association with the response, a large number of RFs (Hapfelmeier and Ulm (2013) used 400 in their studies) has to be computed. Each RF is constructed based on a different permuted version of the variable and the importance score of the permuted version is computed. The p-value for the variable is then computed as the fraction of variable importance scores (obtained for the permuted versions), that are greater than the variable importance of the original (i.e., unpermuted) version of the variable. The computation of p-values for all variables thus requires computing as many RFs as predictor variables multiplied by the number of permutation runs. This approach has been developed and investigated for the low-dimensional setting which typically includes not more than a dozen covariates. It is obvious that with high-dimensional data such permutation-based approaches become very computationally demanding, and might even become practically unfeasible.

A heuristic variable importance test for high-dimensional data is presented in this chapter that is computationally very fast and particularly suitable for high-dimensional genomic data. This test is based on a slightly modified version of the permutation VIM. Note that the permutation VIM is the method of first choice for a VIM since it is almost unbiased. In contrast to the existing approaches, the novel testing procedure is not based on permutations. The idea of this novel testing procedure is to use the information of observed non-positive variable importance scores to reconstruct the null distribution of variable importance. This null distribution is then used to compute p-values. Results of several studies are shown that explore if the new testing approach controls the type I error and investigate its power in settings with binary response. The power of the novel testing approach is also compared to the power of the permutation-based testing approach of Altmann et al. (2010). The approach of Altmann et al. (2010) has often been used since its introduction in 2010 (e.g., Polak et al.; 2015; Prosperi et al.; 2014). It is very computationally demanding, especially for high-dimensional data settings. But compared to the approach of Hapfelmeier and Ulm (2013) it is computationally feasible for high-dimensional data settings. Therefore only the testing approach of Altmann et al. (2010) is considered as a competing method.

4.2. Methods

In the first part of this section, existing and new testing procedures are described which can be used to select relevant variables. A variable is termed as relevant if the trees' prediction errors significantly increase after the random permutation, or equivalently, if the variable significantly improves the prediction accuracy. It is important to note that this definition of relevant predictor variables also includes variables that do not have their "own" effect on the response, but are associated with the response due to their correlation with truly influential predictor variables. From the definition of the VIM, it is clear that negative values or values of zero indicate that the variable does not improve the trees' predictive abilities, because on average the error rates are similar or even larger when using the original, that is, unpermuted version of the variable. Thus it is concluded that the variable is likely to not be relevant. A positive value for the variable importance, in contrast, reflects that the variable at least slightly improves the trees' predictive abilities since the error rates are smaller on average when using the original version of the variable for deriving tree predictions. However, one cannot infer that a positive value for the variable importance indicates a relevant variable since one does not know if the change in prediction errors is solely due to chance. Testing procedures are required to assess if the change in error rates is significantly larger than zero. If it is the case one can infer that the variable is likely relevant.

4.2.1. Permutation-based testing approach of Altmann et al. (2010)

The testing approach of Altmann et al. (2010) has originally been proposed as heuristic for correcting biased VIMs, such as the Gini VIM. However, it is applicable to all kinds of VIMs of RF. Besides its ability to correct biased VIMs, it outputs p-values which are computed from importance scores. This feature enables the user to select relevant variables based on the p-values.

In the first step of the method of Altmann et al. (2010), the variable importance scores are obtained for all variables. Any arbitrary VIM may be used for computing the importance scores – it may even be biased. In the second step, importance scores for settings in which the variable is not associated with the response are computed. Altmann et al. (2010) generate these settings by randomly permuting the response variable to break any associations between the response variable and all predictor variables. The data generated in this way is then used to construct a new RF and to compute the importance scores for the predictor variables. The importance scores can be regarded as realizations drawn from the unknown null distribution. The procedure, which involves the steps of randomly permuting the response vector, constructing a RF and computing the importance scores,

is repeated S times. For each variable there are S importance scores that can be regarded as realizations from the unknown null distribution. Finally, in the last step of the method of Altmann et al. (2010), the S importance scores are used to compute the p-value for the variable. One possibility for deriving the p-value consists in computing the fraction of S importance scores that are greater than the original importance score. This approach is referred to as the *non-parametric* approach since no assumptions are made on the distribution of importance scores of unrelated predictor variables. Alternatively, one can assume a parametric distribution such as the Gaussian, Log-normal or Gamma-distribution for the importance scores of unrelated predictor variables. The parameters for the respective distribution are replaced by their maximum likelihood estimates, which are computed based on the S importance scores of the considered variable. Having defined a specific distribution for the variable's null importance, the p-value is computed as the probability of observing an importance score that is higher than the original importance score, given this distribution. This approach is referred to as *parametric* approach.

4.2.2. Naive testing approach

From its definition, the importance scores computed based on the classical permutation VIM (Eq. (2.7)) are expected to randomly vary around the value zero if variables are not associated with the response. A new heuristic approach is investigated which consists in approximating the null distribution based on the observed non-positive importance scores. More precisely, the variable importance null distribution is reconstructed by mirroring the empirical distribution of the observed negative and zero importance scores on the y-axis. This results in a distribution which is symmetric around zero (see Figure 4.1). Let $M_1 = \{VI_j | VI_j < 0; j = 1, \ldots, p\}$ denote the observed negative variable importance scores, and $M_2 = \{VI_j | VI_j = 0; j = 1, \ldots, p\}$ is the set of importance scores which are zero, with p denoting the number of candidate predictors. The hypothetical importance scores $M_3 = \{-VI_j | VI_j < 0; j = 1, \ldots, p\} = -M_1$ are defined, which arise from multiplying the negative importance scores by -1. The null distribution $\hat{F}_0$ is computed as the empirical cumulative distribution function of $M = M_1 \cup M_2 \cup M_3$. Based on $\hat{F}_0$ a p-value for variable X_j is derived as

$$p_j = 1 - \hat{F}_0(VI_j).$$

It is clear that this testing approach is not suitable for all types of data. The data must contain a large number of variables without any effect so that the approximation of the null distribution is precise enough. A high number of variables without any effect is typically present with genetic data, such as microarray or SNP data, so that the novel testing approach is primarily of relevance to high-dimensional genomic data settings.

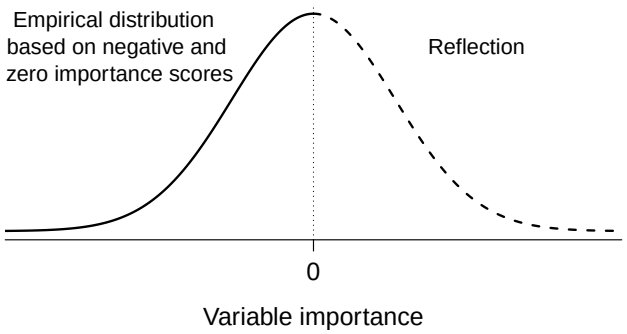

Figure 4.1.: Reconstruction of the null distribution based on variables that are likely non-relevant (i.e., with negative or zero importance scores). The negative part of the null distribution (solid line) is approximated based on the observed negative and zero importance scores. The positive part (dashed line) is obtained from reflection about the y-axis.

4.2.3. Novel variable importance measure based on cross-validation

The novel VIM is not based on the out-of-bag observations but uses a similar strategy, which is inspired by the cross-validation procedure. In brief the idea is as follows: The data is first split into k sets of equal size. Then k RFs are constructed, where the l-th RF is constructed based on observations that are not part of the l-th set. For each RF observations are used for variable importance computation that were not used for constructing the RF.

Let S_l contain the indices of observations from the l-th set, and $|S_l|$ denotes the cardinality of S_l. For categorical response the *fold-specific variable importance* for predictor variable X_j is defined by

$$VI_j^{CV(l)} = \frac{1}{ntree} \sum_{t=1}^{ntree} \frac{1}{|S_l|} \sum_{i \in S_l} \{I(y_i \neq \hat{y}_{it}^*) - I(y_i \neq \hat{y}_{it})\}, \tag{4.1}$$

with $ntree$ denoting the number of trees in a RF, $I(\cdot)$ denoting the indicator function and $\hat{y}_{it}$ and $\hat{y}_{it}^*$ denoting the predictions by the t-th tree before and after permuting the values of X_j, respectively. Note that the predictions $\hat{y}_{it}$ and $\hat{y}_{it}^*, t = 1, \ldots, ntree$, are obtained from the RF, which is constructed based on observations $\{1, 2, \ldots, n\} \setminus S_l$, and thus does not use the observations $i \in S_l$ in tree construction.

The *cross-validated variable importance* for predictor variable X_j is then defined by

$$VI_j^{CV} = \frac{1}{k} \sum_{l=1}^{k} VI_j^{CV(l)}. \tag{4.2}$$

The most simple version of cross-validation results for $k = 2$, so that each of the two sets is once used for creating the RF and once for deriving importance scores. In general, this

method is also known as 2-fold cross-validation or the hold-out method. To differentiate it from cross-validation with $k \geq 3$, from now on it will be referred to as the hold-out method. The corresponding *hold-out variable importance* for variable X_j is given by

$$VI_j^{HO} = \frac{1}{2} \sum_{l=1}^{2} VI_j^{CV(l)}, \tag{4.3}$$

and directly results from setting k to 2 in Eq. (4.2). Thus it is a special case of the cross-validated VIM defined in Eq. (4.2).

4.2.4. Novel testing approach

The new testing approach solely differs from the naive testing approach in the fact that it uses the hold-out VIM (Eq. (4.3)) instead of the classical out-of-bag-based VIM (Eq. (2.7)). The hold-out VIM is preferred over the classical VIM in the new testing approach because it has desirable properties as will be shown. Based on the hold-out VIM, the p-values are derived in exactly the same manner as for the naive approach. The basic steps of the novel testing approach are sketched in the following.

A novel variable importance test for high-dimensional data

Step 1 The data is randomly partitioned into two sets of equal size. Each set is used to create a RF. The two RFs are used to compute the hold-out variable importance VI_j^{HO} (see Eq. (4.3)) for variables $X_j, j = 1, \ldots, p$.

Step 2 The null distribution of the hold-out variable importance is approximated based on the observed non-positive importance scores. For this purpose the following sets are defined:

$M_1 = \{VI_j^{HO} | VI_j^{HO} < 0; j = 1, \ldots, p\}$ (i.e., all negative importance scores),

$M_2 = \{VI_j^{HO} | VI_j^{HO} = 0; j = 1, \ldots, p\}$ (i.e., all importance scores of zero) and

$M_3 = \{-VI_j^{HO} | VI_j^{HO} < 0; j = 1, \ldots, p\} = -M_1$ (i.e., all negative importance scores multiplied by -1),

and the empirical cumulative distribution function $\hat{F}_0$ of $M = M_1 \cup M_2 \cup M_3$ is considered.

Step 3 The p-value corresponding to the variable importance score of predictor variable X_j is computed as

$$p_j = 1 - \hat{F}_0(VI_j^{HO}).$$

Note that the hold-out version of the classical permutation VIM is used, which uses the difference in error rates before and after randomly permuting the values of the considered variable. The proposed testing procedure is very general in the sense that hold-out versions of different permutation-based VIMs might be used, such as the conditional permutation VIM of Strobl et al. (2008), the AUC-based permutation VIM of Janitza et al. (2013), or the VIMs for ordinal responses considered in Chapter 3. It is important to note that if one wants to use a different measure, say, the conditional importance of Strobl et al. (2008), the hold-out version of this measure should be computed, that is, the variable importance should be computed using the splitting procedure described in Section 4.2.3.

The new testing approach is implemented in the R package vita, which is based on the R package randomForest (Liaw and Wiener; 2002). Currently, only the hold-out version of the classical VIM is implemented. The R package vita also contains an implementation of the testing approach of Altmann et al. (2010).

4.2.5. Simulation studies

Since the new testing approach is suitable for high-dimensional genomic data, only settings with large numbers of predictor variables and high signal-to-noise ratios are considered. There is common consensus in the literature that it is very difficult – if not impossible – to simulate realistic complex data structures which capture all the patterns and sources of variability that are generated by a real biological system. Therefore the studies are based on five high-dimensional genomic data sets from real world applications (see Table 4.1 for an overview). These data sets were often used by various authors for binary classification purposes (e.g., Díaz-Uriarte and De Andres; 2006; Dettling and Bühlmann; 2003; Tan and Gilbert; 2003). Note that no pre-selection of data sets based on the results was done, instead the results of all data sets that were analyzed are reported, as has been recommended by Boulesteix (2015).

Data	No. predictors p	No. observations n	Source
Prostate Cancer	6033	102	Singh et al. (2002)
Breast Cancer	4869	77	van't Veer et al. (2002)
Leukemia	7129	72	Golub et al. (1999)
Colon Cancer	2000	62	Alon et al. (1999)
Embryonal Tumor	7129	60	Pomeroy et al. (2002)

Table 4.1.: Overview of high-dimensional genomic data sets used for the investigations.

To study the properties of the novel test, one has to know which of the variables are truly relevant and which are not. In other words, one has to know the truth, which we can never know from real world data. Therefore the design matrix of the real world data sets was used, but the response vector was generated anew according to a specified relation.

Three different studies were performed. Table 4.2 gives an overview of the three studies. In Study I none of the predictor variables of a data set has an effect on the response and there are correlations between predictor variables. In Studies II and III some of the predictor variables have an effect. While Study II includes correlated variables, in Study III all predictor variables are independent of each other.

	Predictor variables with effect	Correlations between predictor variables
Study I	no	yes
Study II	yes	yes
Study III	yes	no

Table 4.2.: Overview of performed studies which differ in the inclusion of predictor variables with effect and in the presence of correlations between predictor variables.

All three described testing procedures were applied in Studies I, II and III. A variable was selected by a testing procedure if its p-value was below $\alpha = 0.05$. To obtain stable results the computations were performed for 500 repetitions of each study. Due to computational reasons, only 200 repetitions of each study were performed for the approach of Altmann et al. (2010). The permutation VIM defined in Eq. (2.7) was used for computing p-values according to the approach of Altmann et al. (2010). This enables a fair comparison of the approach of Altmann et al. (2010) and the novel approach, which is based on the permutation VIM. The p-values for both approaches (non-parametric and parametric) were always computed. Altmann et al. (2010) point out that a Kolmogorov-Smirnov test might be used to choose the most appropriate distribution for the parametric approach. In the present studies, algorithm 1 (outlined in the Supplement to Altmann et al.; 2010) was adhered to, which uses a Gaussian distribution with mean and variance estimated by the arithmetic mean and sample variance, respectively. The parameter S should be chosen so that it is large enough. For the parametric approach the recommendation of Altmann et al. (2010) is a value S between 50 and 100. No recommendations were given for the non-parametric method. A large value $S = 500$ was always used in the studies to exclude the possibility that the performance of Altmann's approach may be related to a suboptimal choice of parameters. In the following each study is described in more detail.

Study I The first study reflects scenarios where all predictor variables are pure noise. The original design matrix and the original response vector of the real data applications were used. To destroy associations between the response vector and the design matrix the elements of the response vector were permuted. In this modified data, associations between predictor variables and the response are only due to chance. Note that the design matrix was not modified and correlations between predictor variables were preserved.

Study II In the second study a scenario was simulated in which 100 variables have an effect on the response and the other variables have no effect. The original design matrix reflecting realistic correlation patterns was again used, but the response vector was generated anew. This allows for a complex data scenario, but at the same time one knows which of the variables are relevant.

The binary response Y for an observation with covariate vector $\boldsymbol{x}^\top = (x_1, \ldots, x_p)$ was generated from a logistic regression model with success probability

$$P(Y = 1|\boldsymbol{x}) = \frac{\exp(x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p)}{1 + \exp(x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p)},$$

with p denoting the total number of predictor variables in the considered data set. The coefficients $\beta_1, \ldots, \beta_p$ were chosen as follows: First $j_1, \ldots, j_{100}$ were randomly drawn without replacement from the set $\{1, \ldots, p\}$ to define which of the variables have an effect on the response and should therefore be selected by a variable importance testing procedure. The corresponding coefficients $\beta_{j_1}, \ldots, \beta_{j_{100}}$ were subsequently drawn from the set $\{-3, -2, -1, -0.5, 0.5, 1, 2, 3\}$, while ensuring that all elements contained in the set are drawn equally often. All other coefficients were set to zero.

Although standardization is not necessary for the application of RF in general, the design matrix was standardized before generating the response in order to make effects comparable across variables of different scales.

Study III This study includes only uncorrelated predictor variables. The design matrix of the real data sets was used and the values within each variable were permuted independently to create uncorrelated variables. As with Study II, 100 variables were supposed to have an effect on the response. The approach for deciding which variables have an effect and for generating the response is exactly the same as described for Study II.

Parameter settings

Analyses under different parameter settings were performed to see if the choice of parameters affects the results. All studies (Studies I, II, III) were performed

- for two different values for the parameter *mtry*: $mtry = \sqrt{p}$ and $mtry = \frac{p}{5}$, with p denoting the total number of predictor variables.

- for two different total numbers of predictor variables. Either a very large number of candidate predictors was used, namely that from the original design matrices (see Table 4.1), or a subset of $p = 100$ predictor variables randomly drawn from the original design matrices. In the studies with large predictor numbers 100 variables

had an effect, and in the studies with a subset of $p = 100$ predictor variables only 20 variables had an effect (only relevant to Study II and III).

- for two different sets that both determine the effects of relevant predictor variables (only relevant to Study II and III). One set was chosen as $\{-3, -2, -1, -0.5, 0.5, 1, 2, 3\}$. The other effect set contained smaller effects; this was chosen as $\{-1, -0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8, 1\}$. Since results were very similar for the two different sets, only those for the effect set $\{-3, -2, -1, -0.5, 0.5, 1, 2, 3\}$ are shown.

Random forest parameter setting

The classical RF version of Breiman (2001) implemented in the R package randomForest (Liaw and Wiener; 2002) was used for the studies shown in this chapter. Though this version implements a biased split selection, it was chosen for implementing the studies because of its computational speed. With respect to computing time, the RF implementation of Breiman (2001) by far outperforms the (unbiased) RF implementation of Hothorn, Hornik and Zeileis (2006). Since settings with a very large number of predictor variables are considered and RF were repeatedly fit, the unbiased RF version of Hothorn, Hornik and Zeileis (2006) is not applicable due to its high computational effort. However, only data settings with continuous predictor variables were chosen to avoid affecting the results by the biased split selection. Thus it is not expected that a split selection bias would occur in the studies. Moreover, subsampling was used instead of bootstrapping in order to avoid possible biases induced by the bootstrap (Strobl et al.; 2007). Subsamples were of size $\lceil 0.632n \rceil$, with n denoting the total number of observations (Strobl et al.; 2007). The number of trees in the RF was always set to 5000. All other parameters not mentioned here were set to the default values so that trees were grown to maximal depth.

Evaluation criteria

One important aspect that was investigated in the studies is the statistical power of the testing approaches. The statistical power is generally defined as the probability of rejecting the null hypothesis, given that the null hypothesis is false. In the current context the null hypothesis states that the trees' prediction accuracy does not worsen when permuting the values of a predictor variable. If the null hypothesis is rejected (i.e., prediction accuracy worsens), there is evidence that the variable is relevant. The statistical power of the testing approaches was explored by computing the fraction of selected variables (i.e., variables with p-value below $\alpha = 0.05$) of those that have an effect. Note that in Studies II and III there are predictor variables with different effect strengths; the absolute effect strengths are $0.5, 1, 2, 3$, or $0.2, 0.4, 0.6, 0.8, 1$ in the alternative setting. For power consider-

ations the proportion of selected variables among variables with the same absolute effect was computed.

The second important aspect concerns the validity of the testing approaches. The type I error of a test is defined as the probability of rejecting the null hypothesis, given that the null hypothesis is true. A test is valid if its type I error does not exceed the significance level α. In the studies it was investigated if the testing procedures control the type I error by computing the fraction of variables with p-value below $\alpha = 0.05$ among those variables that are not relevant. For this purpose one has to know which variables are not relevant. In Study I none of the variables has an effect and thus none is relevant. In Study III exactly those variables whose regression coefficient is zero are not relevant. In Study II, however, due to the correlation between the variables, it is difficult to assess which variables are not relevant: Predictor variables that do not have an "own" effect (i.e., those with coefficient of zero) but are correlated with variables that have an effect, might significantly improve the trees' predictive abilities. Therefore in Study II, the regression coefficients cannot be used to judge which variables are not relevant, because variables with coefficients of zero can also be relevant. Thus only Study I and III can be used for investigating the type I error.

In addition to type I error and power investigations, two further related issues were inspected. The first issue concerns the assumption of the new testing procedure that under the null hypothesis the variable importance distribution is symmetric around zero. It was empirically assessed if this is the case for the classical VIM (Eq. 2.7) and the novel VIM introduced in Section 4.2.3 by plotting the distribution of variable importance scores observed in Study I, where none of the variables is relevant. If an asymmetric distribution or a distribution which is shifted along the x-axis is observed, the testing procedure is expected to have a too high or too low type I error.

The second issue concerns the discrimination between relevant and non-relevant variables by their importance scores. A testing procedure will have low statistical power if it is based on a VIM that does not discriminate well between relevant and non-relevant variables. Thus the discriminative ability was inspected to see if the novel hold-out VIM may be used in a testing procedure. The classical permutation VIM was considered as "gold standard". Its discriminative ability was compared to that of the hold-out VIM. The area under the curve (AUC) was used as a measure for discriminative ability. It here corresponds to an estimate of the probability that a randomly drawn relevant variable has a higher importance score than a randomly drawn non-relevant variable (cf. Section 3.3.2). An AUC value of 1 means that each of the relevant variables receives a higher importance score than any non-relevant variable, thus indicating perfect discrimination by the VIM. An AUC value of 0.5 means that a randomly drawn relevant variable receives a higher

importance score than a randomly drawn non-relevant variable in only half of the cases, indicating no discriminative ability of the VIM.

4.3. Results

4.3.1. Properties of the classical and novel variable importance

Null distribution

Figure 4.2 shows the null variable importance distributions for the novel hold-out VIM and the classical VIM for the settings with large predictor space and *mtry* set to $\sqrt{p}$. Results are very similar for $mtry = \frac{p}{5}$ and are shown in Figure C.9.

The null distribution of the hold-out variable importance seems to be symmetric around zero, and thus seems to satisfy the presumption of a symmetric null distribution. In contrast to that, the null distribution of the classical variable importance is not totally symmetric. In the studies with $p = 100$, this asymmetry is much more apparent (Figure C.15): All distributions are clearly positively skewed showing that a large fraction of variables have small negative importance scores, while smaller fractions of variables have large positive importance scores. The null distribution of the cross-validated variable importance looks very similar for $k \geq 3$ (see Figures C.9, C.15). In contrast, the null distribution of the fold-specific variable importance is nearly symmetric around zero (results not shown). This seems to be contradictory since the cross-validated variable importance is the average of fold-specific variable importances. Further inspection of the simulation results reveals that this effect is possibly due to the overlap of RFs. For $k \geq 3$ the same observations are used for creating the RFs of several folds. For example, in the case of three sets, S_1, S_2, S_3, the first RF is constructed using S_2 and S_3, the second RF is constructed using S_1 and S_3, and the third RF is based on S_1 and S_2. Each pair of RFs have some part of the observations in common. For example, the first and the second RFs are both based on observations from set S_3. The variables have similar predictive abilities for the sets $S_2 \cup S_3$ (on which the first RF is trained) and $S_1 \cup S_3$ (on which the second RF is trained). If high values for a variable X_j speak in favor of class 1 in the subset $S_2 \cup S_3$, then in the subset $S_1 \cup S_3$ high values for X_j will also speak in favor of class 1 – even if there is, in truth, no association between X_j and the class membership. Even in settings without any associations, the two RFs then often select the same predictor variables for a split. Thus for $k \geq 3$ the same few variables will always obtain high fold-specific importance scores, as also seen from empirical studies. In Figure 4.3 the fold-specific variable importance scores for the first two folds (for the Colon Cancer data) are plotted against each other for different values of k. The fold-specific variable importance computed for 500 repetitions

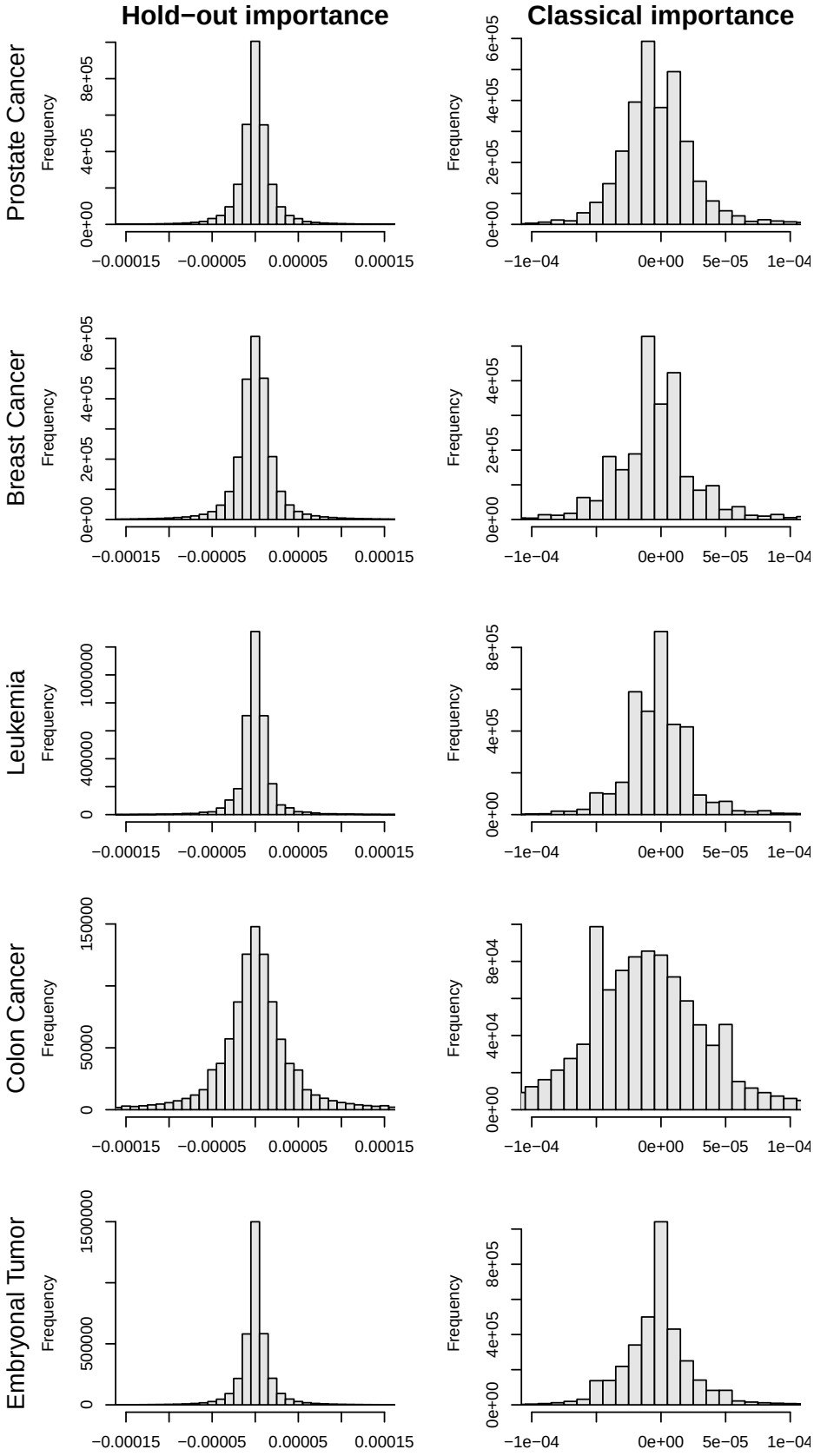

Figure 4.2.: Variable importance null distribution when using the classical VIM and the novel hold-out VIM in settings with large numbers of predictor variables and $mtry = \sqrt{p}$. Distributions are shown for 500 repetitions of Study I.

of Study I (no relevant variables) are shown; similar results are obtained for the other data sets and when using only a subset of $p = 100$ predictor variables (not shown). For $k \geq 3$ the phenomenon just described is clearly observed: There are some variables which have large positive fold-specific variable importance scores for both folds resulting in a large cross-validated variable importance score. In contrast, there are not as many variables with negative fold-specific variable importances for both folds. From that it is clear that the cross-validated variable importance has a skewed null distribution.

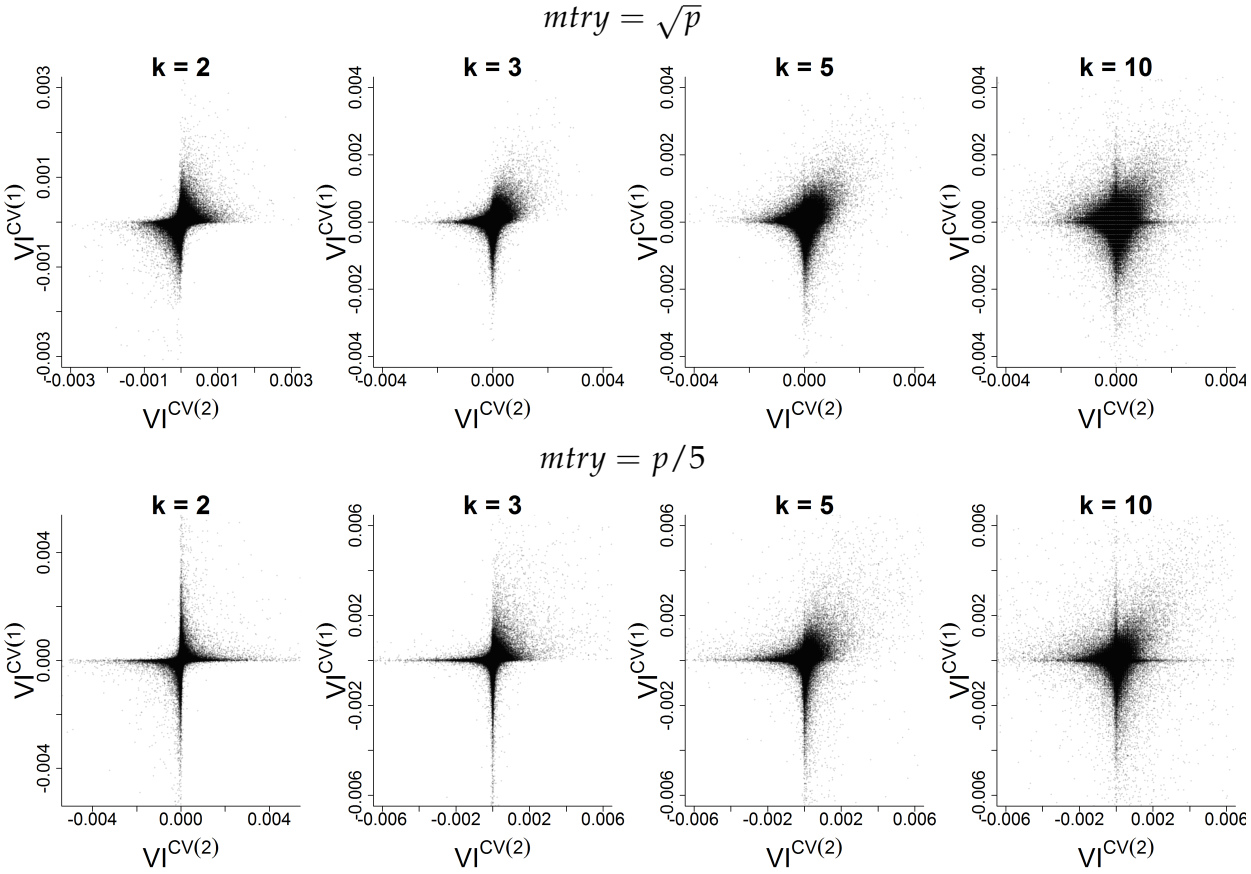

Figure 4.3.: Fold-specific variable importance for the first fold plotted against fold-specific variable importance for the second fold for Study I of the Colon Cancer data with $k = 2$, $k = 3$, $k = 5$ and $k = 10$. Results are shown for settings with large predictor numbers and $mtry = \sqrt{p}$ (upper row) and $mtry = \frac{p}{5}$ (lower row).

It is expected that similar mechanisms occur with the classical VIM which is based on the out-of-bag observations, as the classical VIM is similar to the cross-validated VIM in which k is set to the total number of observations, n. But more research is needed to fully understand the behavior of the classical VIM. The hold-out VIM, in contrast, is not affected in the same manner. Here the data is partitioned into the sets S_1 and S_2. Each set – and correspondingly each observation within the set – is used for the construction of one RF. The first RF uses S_2 and the second RF uses S_1, resulting in two RFs which are completely independent of each other. The selection of variables for a split in the second RF is thus independent of which variables have been selected in the first RF. Therefore

the mechanisms described for $k \geq 3$ do not apply for $k = 2$. This is also supported by the results in Figure 4.3 (first column) where an equal amount of variables with negative fold-specific variable importance scores is observed for both folds as variables with positive fold-specific importance scores for both folds. Although, there is a substantially higher number of variables with both negative or positive fold-specific importance scores than variables with one negative and one positive fold-specific importance score. This might be explained by the fact that the variable importance for the first RF is computed using observations from set S_1, that have been used for the construction of the second RF, and vice versa. A positive correlation might therefore be expected between the fold-specific importance scores. However, this has no effect on the symmetry of the null distribution of the hold-out variable importance.

To conclude, it was empirically shown that the hold-out variable importance has a symmetric null distribution, while the classical importance and the cross-validated variable importance do not have a symmetric distribution. From the studies it is expected that the novel testing approach controls the type I error exactly, while the naive testing approach does not.

Discriminative ability

Figure 4.4 shows the discriminative ability of the classical and the hold-out VIMs for Study II and Study III. Results are shown for the settings with large predictor space and *mtry* set to $\sqrt{p}$. The novel hold-out VIM and the classical VIM had very similar discrimination ability. For Study III the performance of the hold-out VIM was slightly better than the performance of the classical permutation VIM. The results with $mtry = \frac{p}{5}$ are very similar (Figure C.8), and a slightly better performance of the hold-out VIM can be observed in both, Study II and III. The results for the predictor number reduced to $p = 100$ are in line with these findings and are shown in Figures C.17 and C.18. Therefore the novel hold-out VIM is considered a good measure to reflect the relevance of variables. The cross-validated VIM with $k \geq 3$ has similar discriminative ability, too (results not shown). As with the classical VIM, when using the hold-out and cross-validated VIMs each observation is used for tree construction and for variable importance computation. In contrast to that, the fold-specific VIM, defined in Eq. (4.1) uses one part of the observations only for tree construction and the other part for variable importance computation. By building an average of fold-specific importances one makes sure that all information is used for tree construction and for variable importance computation.

To summarize, the studies indicate that the hold-out VIM does not have a worse discriminative ability than the classical VIM and thus might be used as an alternative to the classical VIM. The distribution of hold-out importance scores, in addition, is symmetric

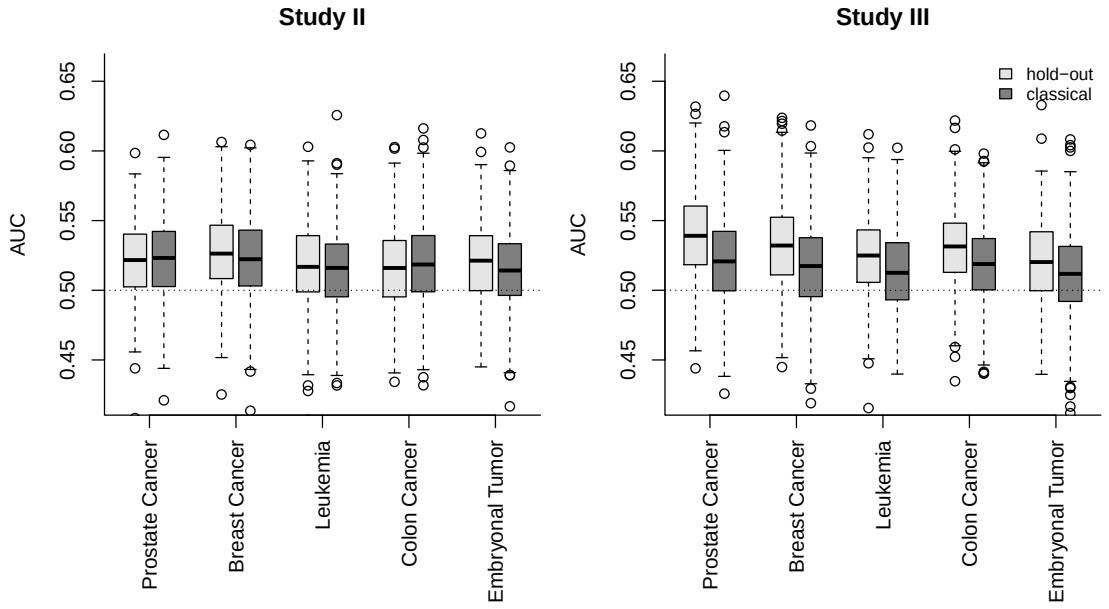

Figure 4.4.: Discriminative ability of the novel hold-out VIM and the classical VIM for Study II (left) and Study III (right) in settings with large predictor numbers and $mtry = \sqrt{p}$. Discriminative ability is measured by the area under the curve (AUC). Values of 0.5 indicate no discriminative ability (horizontal dotted line).

around zero for variables not associated with the response – a criterion that is not fulfilled for the classical and the cross-validated VIMs. This motivates the use of the hold-out VIM in the proposed testing procedure.

4.3.2. Type I error

The type I errors of the three testing approaches were investigated using Study I and are depicted in Figures 4.5 and 4.6. The type I errors of the novel testing procedure were always close to the significance level $\alpha = 0.05$, indicating that the test does not reject the null hypothesis too often or too rare. These findings are in line with the results in Section 4.3.1 where it was shown that the null distribution of the hold-out variable importance is nearly symmetric around zero. The results for the naive approach are also in line with the findings from Section 4.3.1. As expected, the type I error of the naive approach is systematically different from 0.05. In the studies with large predictor numbers, the naive approach always gave slightly too large type I errors if *mtry* was set to the default value $\sqrt{p}$, and too small type I errors if *mtry* was $\frac{p}{5}$ (Figure 4.5). In the studies with the predictor number reduced to $p = 100$, the type I errors were always close to 0.1 for both large and small *mtry* values, as seen in Figure 4.6. Therefore the naive approach should only be used with caution.

The non-parametric approach of Altmann et al. (2010) always gave type I errors close to 0.05 for both the studies with large and smaller ($p = 100$) predictor numbers. The

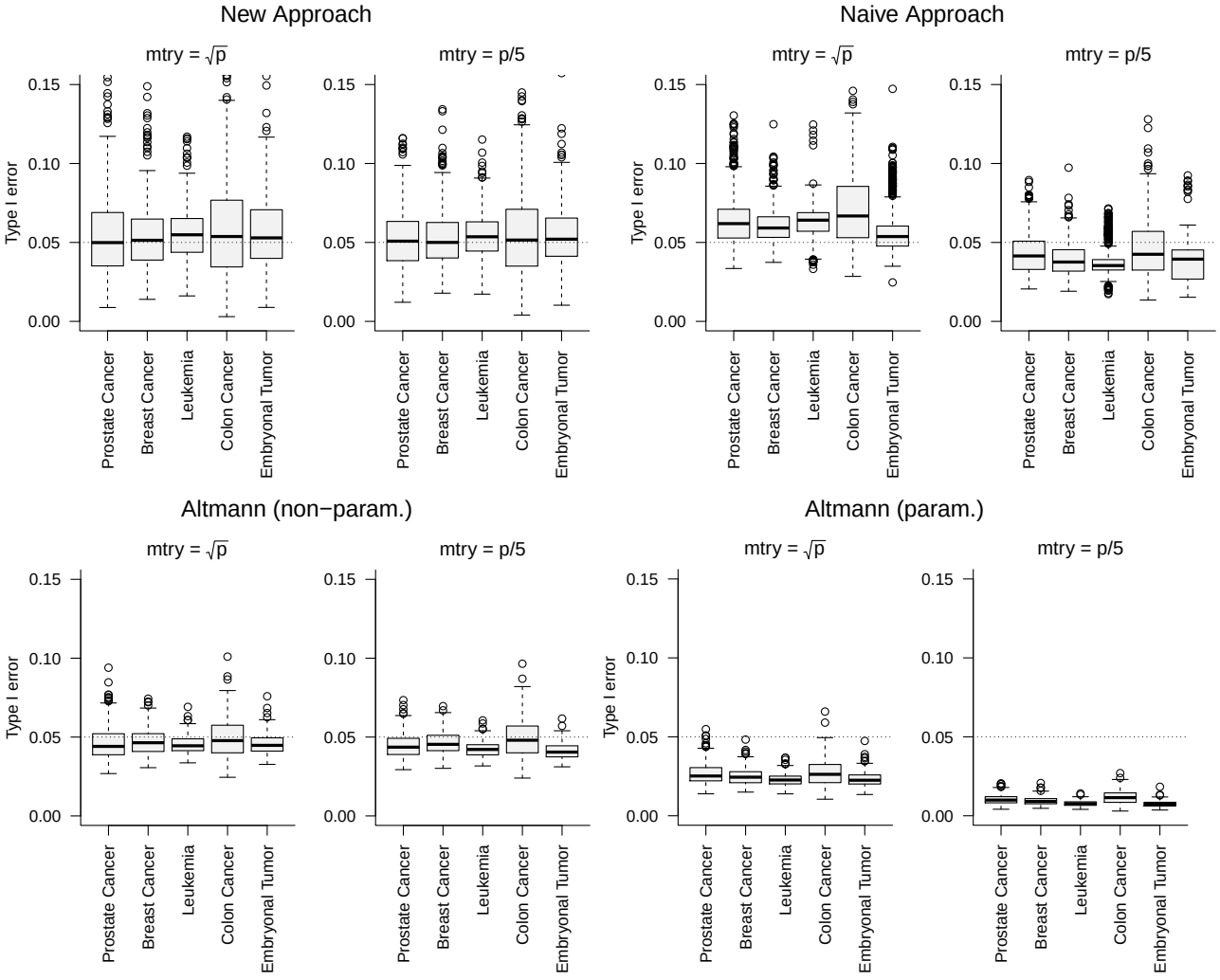

Figure 4.5.: Type I error in Study I for settings with large numbers of predictor variables. Results are shown for the new testing approach, the naive testing approach and the approach of Altmann et al. (2010) (non-parametric and parametric). Hypothesis tests were performed at significance level $\alpha = 0.05$ (dotted horizontal line).

type I error for the parametric approach of Altmann et al. (2010) was always considerably smaller than 0.05 in the studies with large predictor numbers, indicating that the parametric approach is too conservative in settings with large predictor numbers. In the studies with $p = 100$, in contrast, the type I error was much closer to 0.05. The variability in type I errors was smaller for the approach of Altmann et al. (2010) than for the novel and the naive testing procedures. In settings with the predictor number reduced to $p = 100$, the variability was larger for all testing approaches.

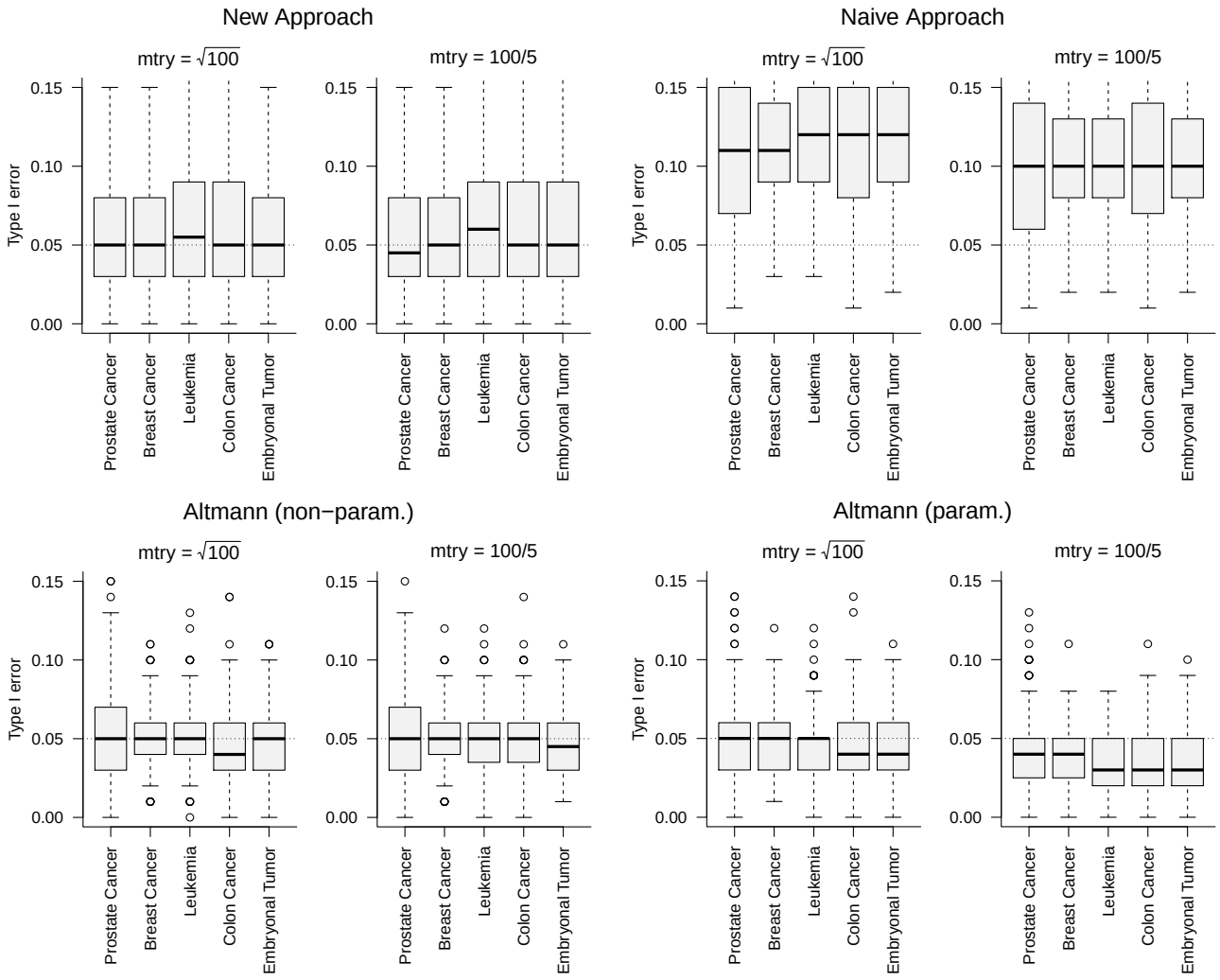

Figure 4.6.: Type I error in Study I for settings with a subset of $p = 100$ predictor variables. Results are shown for the new testing approach, the naive testing approach and the approach of Altmann et al. (2010) (non-parametric and parametric). Hypothesis tests were performed at significance level $\alpha = 0.05$ (dotted horizontal line).

4.3.3. Statistical power

Study III Figure 4.7 shows the proportion of selected variables (i.e., variables with p-value below 0.05) averaged over 500 (200 for the approach of Altmann et al. (2010), resp.) repetitions of Study III for settings with large numbers of variables. The proportions were

computed among variables with the same absolute effect size of 0.5, 1, 2 and 3, respectively. In addition, the proportion of selected variables among variables without effect (i.e., those variables X_j for which $\beta_j = 0$) are shown. For all testing procedures the proportion of selected variables increases with increasing absolute effect size suggesting that variables with larger effects are more easily identified than variables with smaller effects. The parametric approach of Altmann et al. (2010) consistently has the smallest power. The non-parametric approach of Altmann et al. (2010) and the new and the naive testing approaches have similar performance. However, the novel approach has slightly higher statistical power than the non-parametric approach of Altmann et al. (2010), especially in settings with $mtry = \frac{p}{5}$. For $mtry = \sqrt{p}$ a slightly higher number of variables was selected with the naive approach than with the other two approaches for both, non-relevant variables (i.e., variables X_j with $\beta_j = 0$) and relevant variables (X_j with $\beta_j \neq 0$). In contrast, for $mtry = \frac{p}{5}$ fewer variables were selected with the naive approach. The results are in line with the results in Section 4.3.2, where it was shown that the type I error was smallest for the non-parametric approach of Altmann et al. (2010), and was higher (lower) for the naive approach than for the novel approach if *mtry* was set to the default value $\sqrt{p}$ (the value $\frac{p}{5}$).

To conclude, the novel testing approach showed the best performance in the settings with large numbers of variables because it consistently had the highest power while preserving the type I error.

However, the statistical power of all testing procedures was low. In the studies with a subset of $p = 100$ predictor variables, much higher statistical power for all approaches is observed (Figure 4.8). The naive approach did not preserve the type I error in the settings with reduced predictor space. This can be seen when inspecting the proportion of rejections among predictor variables X_j with $\beta_j = 0$ in Figure 4.8. The same was seen from the results of Study I (Figure 4.6). The novel testing approach has similar – and on average even slightly higher – statistical power than the non-parametric and parametric approaches of Altmann et al. (2010).

Note that the results presented so far are averaged over all repetitions of Study III. Thus, there is no information on the variability in the selected number of variables with effect. Further inspection reveals, however, that the variabilities for the naive approach, the novel approach and the non-parametric approach of Altmann et al. (2010) are similar (see Figures C.10 - C.14, C.19 - C.23). The variability for the parametric approach of Altmann et al. (2010), in contrast, was smaller, which is due to the fact that the approach was very conservative and selected only few variables.

Study II The results for Study II with large variable numbers are shown in Figure 4.9. The proportion of selected variables was always largest when using the novel testing ap-

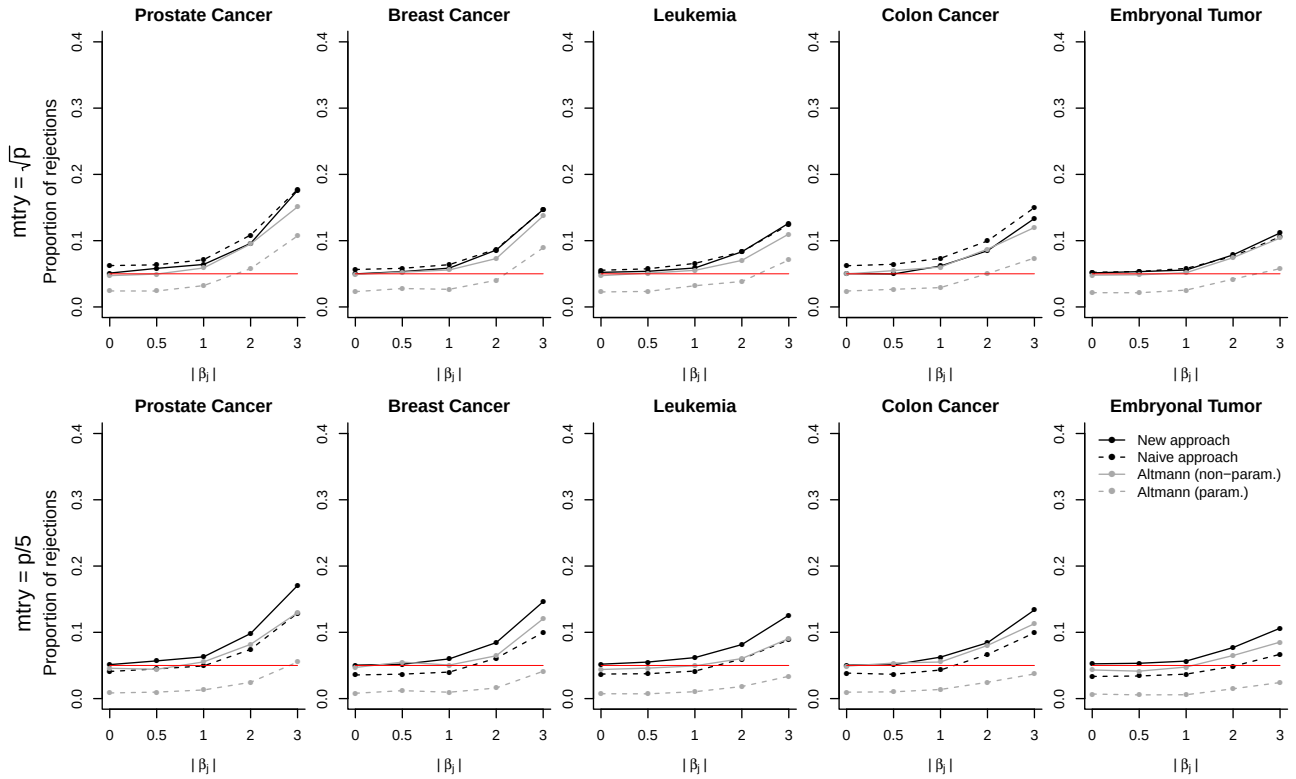

Figure 4.7.: Proportion of rejected null hypothesis among predictor variables X_j with specified absolute effect size $|\beta_j| \in \{0, 0.5, 1, 2, 3\}$ in settings with large predictor numbers. The mean proportions over 500 (200 for the approach of Altmann et al. (2010), resp.) repetitions of Study III are shown when using the novel approach, the naive approach and the approach of Altmann et al. (2010), with *mtry* set to $\sqrt{p}$ (upper panel) and $\frac{p}{5}$ (lower panel). The red horizontal line represents the 5% significance level.

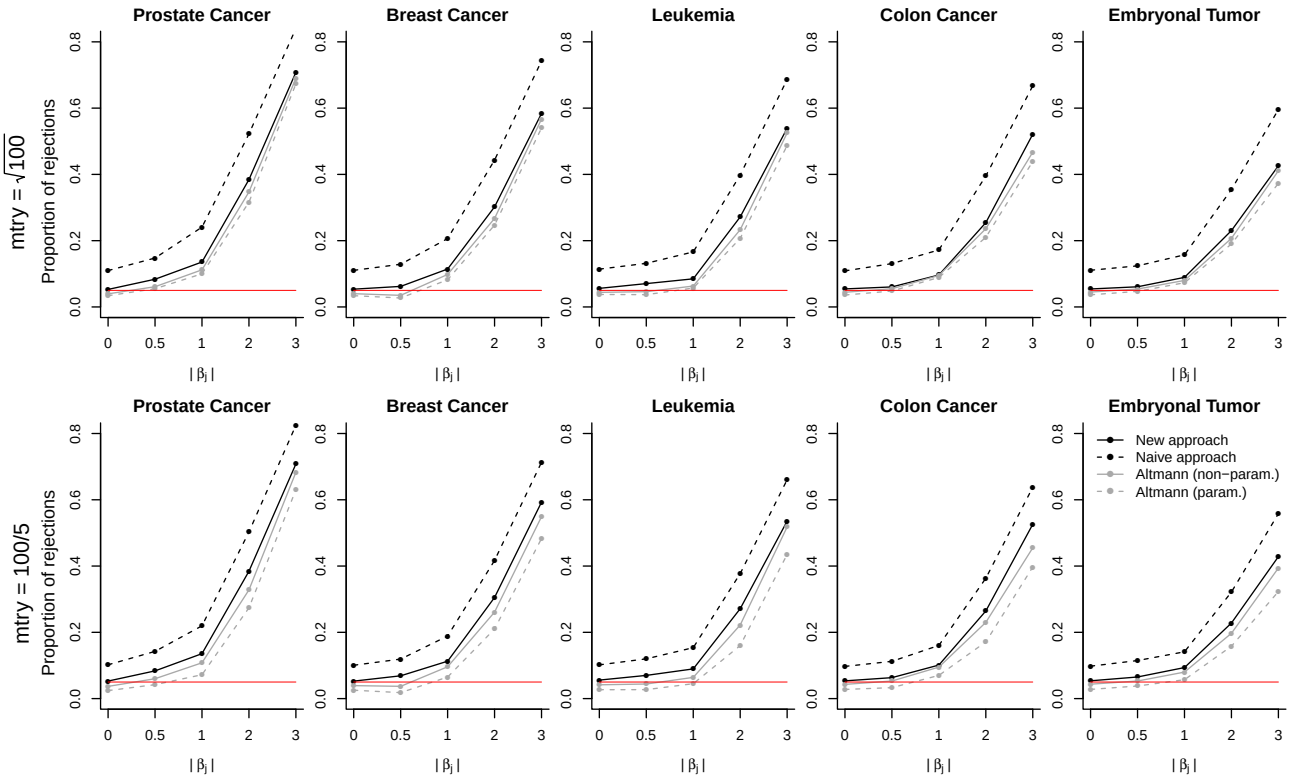

Figure 4.8.: Proportion of rejected null hypothesis among predictor variables X_j with specified absolute effect size $|\beta_j| \in \{0, 0.5, 1, 2, 3\}$ in settings with a subset of $p = 100$ predictor variables. The mean proportions over 500 (200 for the approach of Altmann et al. (2010), resp.) repetitions of Study III are shown when using the novel approach, the naive approach and the approach of Altmann et al. (2010), with *mtry* set to $\sqrt{100}$ (upper panel) and $\frac{100}{5}$ (lower panel). The red horizontal line represents the 5% significance level.

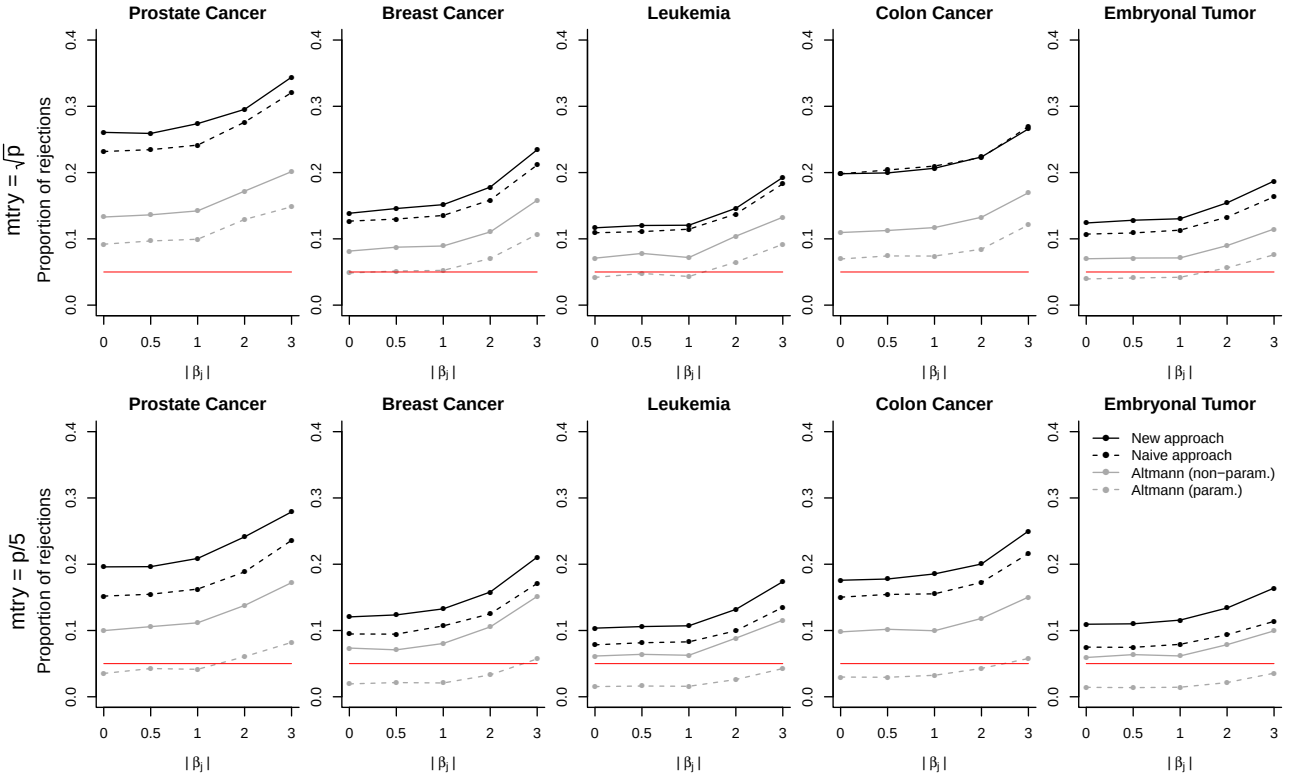

Figure 4.9.: Proportion of rejected null hypothesis among predictor variables X_j with specified absolute effect size $|\beta_j| \in \{0, 0.5, 1, 2, 3\}$ in settings with large predictor numbers. The mean proportions over 500 (200 for the approach of Altmann et al. (2010), resp.) repetitions of Study III are shown when using the novel approach, the naive approach and the approach of Altmann et al. (2010), with *mtry* set to $\sqrt{p}$ (upper panel) and $\frac{p}{5}$ (lower panel). The red horizontal line represents the 5% significance level.

proach. Thereafter, the proportion decreases bit by bit for the naive testing approach, the non-parametric approach of Altmann et al. (2010) and the parametric approach of Altmann et al. (2010). The approaches of Altmann et al. (2010) identified far fewer variables as relevant than the naive and the novel testing procedures. With the parametric approach, the proportion of selected variables was very low, especially if the predictor numbers were large and *mtry* was set to $\sqrt{p}$. It was even lower than 0.05, indicating that the parametric approach of Altmann et al. (2010) is too conservative. This was not the case for the non-parametric approach of Altmann et al. (2010).

The results for settings with $p = 100$ are shown in Figure C.16. As with Study III, more variables were selected by all testing approaches in these settings. In contrast to the results shown in Figure 4.9 for settings with large variable numbers, the proportion of selected variables was always largest when using the naive approach.

Overall, the proportion of identified variables X_j with $\beta_j = 0$ was very large and greatly exceeds 0.05 in many of the settings. This is attributable to the correlations between the variables. From the construction of the naive and the novel testing approach, variables which do not have an "own" effect, but are correlated with variables with effect, may be

considered as relevant as long as they improve the trees' prediction accuracies. Therefore, even variables that do not have a direct influence are very often identified by the naive and novel testing approaches – but still not as often as variables with direct influence. In contrast to that, it is not clear if the approach of Altmann et al. (2010) is also supposed to select variables that do not have a direct influence but correlate with variables that have an effect. Therefore in settings with correlated predictor variables it is not possible to evaluate which testing approach has better performance. When based on the conditional importance of Strobl et al. (2008), the testing procedures would possibly not as often select variables that are only associated with the response through their correlation to truly influential variables.

4.4. Discussion

During the last years, several approaches have been developed for hypothesis testing based on RF's VIMs (see Hapfelmeier and Ulm; 2013, and references therein). The existing approaches are computationally demanding and require the repeated computation of forests. In this chapter, a computationally fast heuristic approach for a variable importance test was presented, that tests if a predictor variable significantly improves the trees' predictive abilities. The new testing procedure is based on a slightly modified version of the permutation VIM, whose null distribution was shown to be symmetric around zero. The classical permutation variable importance, in contrast, has a skewed null distribution and thus seems inappropriate for the application of the testing procedure. The testing approach based on the classical permutation VIM worked quite well for settings with huge predictor numbers, but did not preserve the type I error in settings with fewer ($p = 100$) predictor variables. It should therefore be used with caution. The use of the testing procedure which is based on a modified version of the permutation VIM, is strongly recommend. This approach has consistently been shown to precisely preserve the type I error in the considered studies with categorical response. Moreover, it successfully identified at least as many relevant predictor variables as the testing approach of Altmann et al. (2010).

The novel testing procedure focuses on the identification of predictor variables which significantly improve the trees' predictive abilities. The permutation VIM, by its definition, reflects the improvement in predictive abilities if a variable is used for making the prediction. Thus, there is a monotone relationship between the value of the variable importance and the p-value derived from the new testing approach: predictor variables with higher importance scores obtain smaller p-values. This must not necessarily be the case with permutation-based approaches. This is obvious as Altmann et al. (2010) state

that their approach corrects for the bias in the Gini VIM which ranks, for example, variables with many categories higher than variables with fewer categories. In this case a re-sorting of variables occurs when computing p-values from the Gini importance based on the proposed permutation procedure. If using a parametric function for the null distribution the permutation-based approach of Altmann et al. (2010) was very conservative in the studies presented in this chapter and had much smaller statistical power than the novel approach. When deriving p-values in a non-parametric way, that is without making any distributional assumptions, the testing approach of Altmann et al. (2010) showed almost the same statistical power as the novel approach. This suggests that the poor performance is related to the assumed parametric distribution of the importance scores of unrelated variables. In the studies the normal distribution was used for modeling the variable importance distribution of unrelated variables. Studies indicate that the assumption of a normal distribution is not reasonable due to the skewness of the distribution of null importance scores (data not shown). Researchers who apply the approach of Altmann et al. (2010) to high-dimensional data should therefore consider alternative distributions or approximate the null distribution in a non-parametric way.

Overall, the statistical power of all testing procedures was low in studies including more than 2000 predictor variables. The power of the VIMs to discriminate between relevant and non-relevant variables was poor, too. The approach of Altmann et al. (2010), which showed high statistical power in other studies (Molinaro et al.; 2011; Hapfelmeier and Ulm; 2013), also had very low power. This discrepancy is likely related to the fact that the existing studies included only a few variables, while the studies presented in this chapter are based on several thousands of variables. Molinaro et al. (2011) for example focused on candidate-gene studies and considered only a few dozens of the features. When performing the studies with a subset of 100 variables the statistical power substantially increased, and the VIMs discriminated much better between relevant and non-relevant variables. This suggests that the issue of detecting relevant features by VIMs is much more difficult for genome-wide association studies, including hundreds of thousands to millions of features, than for candidate-gene studies, that include only a few hundreds of features.

The novel testing approach is, however, not applicable to any high-dimensional data set. It is expected that it may perform poorly if only a few non-positive importance scores are observed. If there are only a few variables with negative importance score or importance score of zero, the approximation of the variable importance null distribution might be too imprecise and might lead to inaccurate p-values. In the most extreme setting (100 predictor variables in total and correlations between predictor variables), on average about 70 non-positive importance scores were observed (for the Prostate Cancer

data even only 40). However, the novel approach still worked surprisingly well. Nevertheless, in settings (i) with small predictor numbers (below 200), or (ii) with very strong correlations between predictor variables, or (iii) with high expected signal-to-noise ratio, it is recommended that users look closely at the number of non-positive importance scores. If this number is small, it is recommended that users be careful when applying the novel testing approach because it is not clear if a small number of non-positive importance scores is sufficient to derive p-values. In such cases one should consider the computationally more demanding alternatives, such as the approach by Altmann et al. (2010), which had very similar performance in the studies presented in this chapter.

5. Hypothesis testing on bootstrap samples

Some approaches have been proposed in the biometrical field where hypothesis testing is performed on a bootstrap sample as if it were the original sample. However, the resulting p-values do not represent what would be obtained on the original data. This chapter explores the reasons for this and assesses the practical impact on procedures relevant to biometrical applications. It is mainly based on Janitza, Binder and Boulesteix (2016) but also contains some results which were presented in Rospleszcz et al. (2016), in which I am the joint first co-author. The work of Rospleszcz et al. (2016) is a result of a master thesis which was supervised by myself.

The structure of this chapter is as follows: Section 5.1 outlines the problem which is addressed in this chapter. In Section 5.2 it is shown through theoretical and empirical results that there is increased type I error for both the Z-test and the likelihood ratio (LR) test when using bootstrapped p-values. The distribution of bootstrapped p-values is subsequently explored in this section. In Sections 5.3, 5.4 and 5.5 the consequences for three practices are investigated, namely, bootstrapping p-values for multivariable model building (*Application 1*; Rospleszcz et al. (2016)), for variable ranking (*Application 2*) and for assessing the variability of p-values (*Application 3*).

5.1. Introduction

Bootstrap procedures are becoming more and more widely used, as indicated by the now large number of reference textbooks on the subject (Chernick; 2008; Manly; 2006; Good; 2005; Davison; 1997). Bootstrapped estimates can be used to derive for example the variance of an estimator, a quantile of interest or a confidence interval (Davison; 1997). This chapter deals with cases where a p-value of a standard statistical test (such as, e.g., the Z-test or the LR test) takes the role of the estimator which is being bootstrapped. More precisely, p-values are meant that result from statistical tests performed using a bootstrap sample as the data set as if it were the original data set, ignoring that it has actually been drawn with replacement from another sample. It is important to note that such pro-

cedures are fundamentally different from obtaining p-values by the so-called bootstrap tests (Efron and Tibshirani; 1993). Bootstrap tests are an alternative to inference based on parametric assumptions when these assumptions are questionable or when such a method simply does not exist. Bootstrap tests as well as their pitfalls and some potential solutions have been extensively discussed in the literature in recent decades; see Efron and Tibshirani (1993) for an overview. In this thesis p-values obtained by these bootstrap tests are not referred to when speaking of bootstrapped p-values. Instead p-values are meant that are obtained from performing any statistical test using a bootstrap sample as the data set as if it were the original, which is a completely different approach. Such bootstrapped p-values have been far less investigated than the famous bootstrap tests that are described for example in Efron and Tibshirani (1993). However, procedures based on bootstrapped p-values are not uncommon in the literature, especially in biometrical applications. They have been used in the statistics and bioinformatics literature for investigating the stability of stepwise model selection procedures (Chen and George; 1985; Altman and Andersen; 1989; Sauerbrei and Schumacher; 1992), for ranking genes with respect to their differential expression (Mukherjee et al.; 2003), for estimating the variability of p-values which one would observe when repeating an experiment multiple times (Boos and Stefanski; 2011) or, in a completely different context, for deciding which variable should be selected for splitting in random forests (Hothorn, Hornik and Zeileis; 2006). In all these applications it is essential that quantities such as p-values computed on bootstrap samples represent what would be obtained on the original data or new data drawn from the overall population. Some articles suggest that this might often not be the case (Bollen and Stine; 1992; Strobl et al.; 2007). These handle very specific cases and a simple general theory to explain the problem is lacking. Further, the practical consequences for biometrical applications are to date largely unknown. This chapter addresses these problems. It gives new theoretical insights and investigates the practical consequences of three specific applications proposed in the literature which are based on bootstrapped p-values.

5.2. Bootstrapping p-values

5.2.1. Type I error

This section outlines the computation of p-values based on a bootstrap sample when ignoring that the sample was drawn from the empirical distribution and not from the true distribution, and shows that the type I error of the corresponding tests is increased. The Z-test and the LR test are used as examples.

Z-test

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ be realizations drawn from $N(\mu, \sigma^2)$ and let $\hat{F}$ denote the corresponding empirical distribution with known σ^2. The test statistic for testing the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$ is given by $Z = \sqrt{n}(\bar{x} - \mu_0)/\sigma$, with $\bar{x}$ denoting the sample mean. Then Z follows a normal distribution with $\mathrm{E}(Z) = \sqrt{n}(\mu - \mu_0)/\sigma$ and $\mathrm{Var}(Z) = 1$.

Now let $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)^\top$ denote the realizations of a bootstrap sample that was drawn from $\hat{F}$ with replacement. The bootstrapped test statistic from a Z-test with hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ is defined as

$$Z^* = \sqrt{n}\frac{\bar{x}^* - \mu_0}{\sigma}, \tag{5.1}$$

with $\bar{x}^* = \frac{1}{n}\sum_{i=1}^n x_i^*$. If incorrectly assuming that Z^* follows a standard normal distribution, the corresponding bootstrapped p-value for an observed test statistic Z^* is computed as

$$p^* = 2 \cdot (1 - \Phi(|Z^*|)), \tag{5.2}$$

with Φ denoting the cumulative distribution function of the standard normal distribution. The following theorem is used to show that Z-tests for the test statistic Z^* have increased type I error, or equivalently, that decisions made on bootstrapped p-values, p^*, lead to systematically too many false positive findings.

Theorem 1

Let the bootstrapped test statistic for a Z-test with $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ be defined as in Eq. (5.1). The unconditional expectation of this bootstrapped Z-test statistic Z^* is $\mathrm{E}(Z^*) = \mathrm{E}(Z)$, while the unconditional variance of Z^* is $\mathrm{Var}(Z^*) = 2$.

Proof

The expectation of Z^* is derived as

$$\mathrm{E}(Z^*) = \mathrm{E}(\mathrm{E}(Z^*|\hat{F})) = \mathrm{E}(Z).$$

The variance of Z^* can be split into two parts,

$$\mathrm{Var}(Z^*) = \mathrm{Var}(\mathrm{E}(Z^*|\hat{F})) + \mathrm{E}(\mathrm{Var}(Z^*|\hat{F})). \tag{5.3}$$

The first term reduces to

$$\mathrm{Var}(\mathrm{E}(Z^*|\hat{F})) = \mathrm{Var}(Z) = 1. \tag{5.4}$$

Hypotheses	Sign. threshold	Type I error for Z	Type I error for Z^*
$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$ (two-sided test)	$z_{0.95} = 1.64$	0.10	0.24
	$z_{0.975} = 1.96$	0.05	0.17
	$z_{0.995} = 2.58$	0.01	0.07
$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$ (one-sided test)	$z_{0.90} = 1.28$	0.10	0.18
	$z_{0.95} = 1.64$	0.05	0.12
	$z_{0.99} = 2.33$	0.01	0.05

Table 5.1.: Type I error when performing two-sided and one-sided upper Z-tests with predefined significance thresholds for test statistic $Z = \sqrt{n}(\frac{1}{n}\sum X_i - \mu_0)/\sigma$ with $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and for a bootstrapped test statistic $Z^* = \sqrt{n}(\frac{1}{n}\sum X_i^* - \mu_0)/\sigma$.

As far as the second term in (5.3) is concerned, the basic assumption underlying bootstrap estimation of the variance, which can be easily shown in the present simple case (Davison; 1997), is that $\mathrm{Var}(Z^*|\hat{F})$ approximates $\mathrm{Var}(Z)$. Using this result one obtains for the second term

$$\mathrm{E}(\mathrm{Var}(Z^*|\hat{F})) = \mathrm{E}(\mathrm{Var}(Z)) = 1. \tag{5.5}$$

Summing (5.4) and (5.5), Eq. (5.3) results in $\mathrm{Var}(Z^*) = 2$.

□

According to Theorem 1, the unconditional variance of the bootstrapped statistic Z^* is twice the variance of Z. Thus under the null hypothesis that $H_0 : \mu = \mu_0$ (or $H_0 : \mu \leq \mu_0$; $H_0 : \mu \geq \mu_0$ for one-sided tests), the marginal distribution of the bootstrapped statistic Z^* is not the standard normal distribution (see also Appendix C.3 for empirical results). Using the significance threshold $z_{1-\frac{\alpha}{2}}$, the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution, the type I error is $2 \cdot (1 - \Phi(\frac{1}{\sqrt{2}} z_{1-\frac{\alpha}{2}}))$, where Φ is the standard normal distribution function. For a one-sided lower (upper) test with null hypothesis $H_0 : \mu \geq \mu_0$ ($H_0 : \mu \leq \mu_0$), the significance threshold z_α ($z_{1-\alpha}$) is used and the type I error is $\Phi(\frac{1}{\sqrt{2}} z_\alpha)$ (and $1 - \Phi(\frac{1}{\sqrt{2}} z_{1-\alpha})$, respectively). Table 5.1 shows examples for the type I error when performing Z-tests for test statistics Z and Z^*. It can be seen that the type I error is substantially increased when performing Z-tests on bootstrap samples as if they were the original samples.

Likelihood Ratio test

The likelihood ratio (LR) test is used for example when comparing the fit of two nested models, where one model contains restrictions that are not imposed in the other. The likelihood of the restricted model, called the submodel in the following, is termed L_0, while L_1 corresponds to the likelihood of the unrestricted model. The test statistic for the

LR test is defined as twice the difference in log-likelihoods:

$$T = -2(\log(L_0) - \log(L_1)). \tag{5.6}$$

The test statistic T asymptotically follows a non-central χ^2-distribution with df degrees of freedom, which is calculated as the difference in degrees of freedom of the two models, and with non-centrality parameter κ. The asymptotic expectation of the test statistic is given by $\text{AE}(T) = df + \kappa$ and the asymptotic variance is $\text{AVar}(T) = 2df + 4\kappa$. Under the null hypothesis which states that the submodel is true, the non-centrality parameter is zero and thus T asymptotically follows a central $\chi^2(df)$-distribution and has asymptotic expectation $\text{AE}(T) = df$ and asymptotic variance $\text{AVar}(T) = 2df$.

The corresponding bootstrapped test statistic for the LR test is

$$T^* = -2(\log(L_0^*) - \log(L_1^*)), \tag{5.7}$$

with L_0^* and L_1^* denoting the likelihoods for the submodel and the unrestricted model, respectively, both evaluated on a bootstrap sample. The bootstrapped p-value for an observed T^* is defined as

$$p^* = P(\Lambda \geq T^* | H_0), \tag{5.8}$$

with $\Lambda \sim \chi^2(df)$.

Bollen and Stine (1992) gave an approximation for the unconditional asymptotic expectation of the test statistic T^*. They report it as being twice as large as the asymptotic expectation of T in the original sample. They also report the unconditional asymptotic variance of T^* to be larger than the asymptotic variance of T. However, their derivations seem to lack in theoretical foundations, since it is not clear that the asymptotic conditional variance of T^* equals $2df + 4T$. Empirical results shown in Janitza, Binder and Boulesteix (2016) (for the LR test with 1 degree of freedom) and Rospleszcz et al. (2016) (for the LR test with varying degrees of freedom) are in line with the theoretical results of Bollen and Stine (1992). The empirical results of Rospleszcz et al. (2016) are shown in Figure 5.1; the LR test statistic was computed based on 10000 bootstrap samples for the comparison of two nested models: the intercept model and the model including a categorical predictor variable – not associated with the response – with 2, 3, 4, 5, 6 and 7 categories. Each bootstrap sample was generated from a different original sample including $n = 1000$ observations. The results of Rospleszcz et al. (2016) in Figure 5.1 clearly show a discrepancy between the distribution of T and that of T^*. The probability mass in the tail of the distribution of T^* is larger than that of T, which leads to increased type I error for LR tests performed on bootstrap samples. Moreover, the discrepancy between the distributions

for T and T^* is greater for tests with more degrees of freedom. This leads to type I errors which are more increased for tests with larger degrees of freedom. Note that, in contrast to the Z-test, there is no straightforward derivation of the type I error for the LR test because the marginal distribution of T^* is unknown. Therefore empirical results are shown in the following. Table 5.2 presented in Rospleszcz et al. (2016) shows the empirical type I errors for LR tests performed with significance thresholds $\chi^2_{df,0.95}$ (i.e., the 95% quantile of the χ^2-distribution with df degrees of freedom) for bootstrapped test statistics T^*. The empirical type I errors are also shown when using the test statistics T. There is a large increase in type I error in the empirical studies, especially for LR tests with many degrees of freedom. While for 1 degree of freedom the type I error is increased by factor 3, for 6 degrees of freedom it is increased by factor 8.

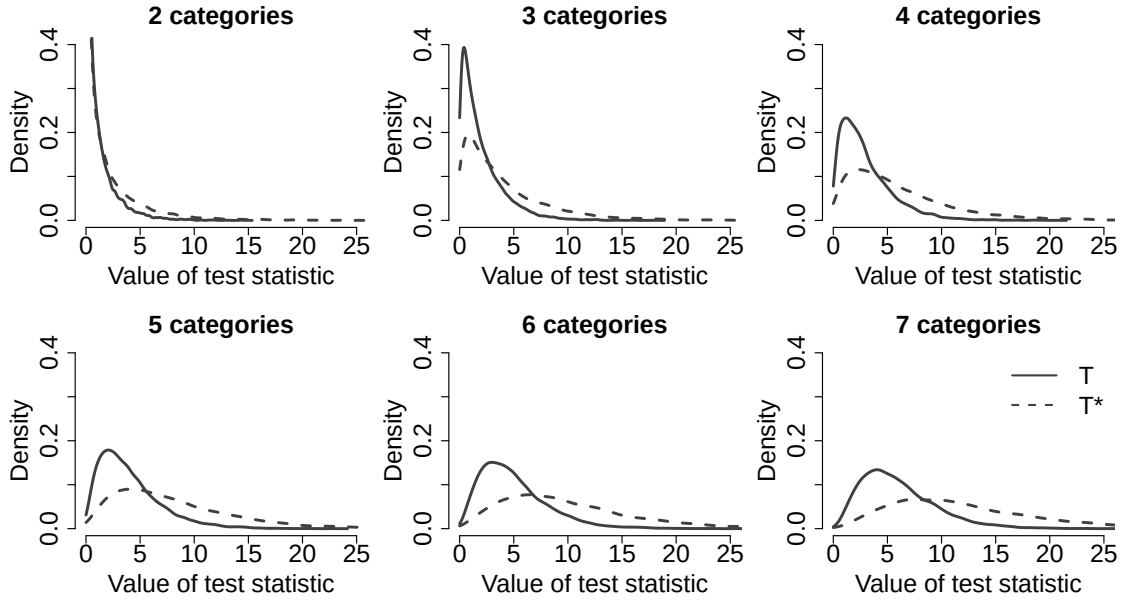

Figure 5.1.: Empirical density functions for LR test statistics T (solid lines) and T^* (dashed lines) for the comparison of the linear regression model with only the intercept to the model including the intercept and one categorical predictor variable with 2, 3, 4, 5, 6 and 7 categories (Rospleszcz et al.; 2016).

		$df = 1$	$df = 2$	$df = 3$	$df = 4$	$df = 5$	$df = 6$
$\chi^2_{df,0.95}$		3.84	5.99	7.81	9.49	11.07	12.59
Empirical type I error	for T	0.055	0.051	0.0483	0.052	0.049	0.049
	for T^*	0.169	0.223	0.272	0.320	0.353	0.393

Table 5.2.: Empirical type I errors when performing LR tests with significance thresholds $\chi^2_{df,0.95}$ for test statistics T and T^*. Bootstrapped and original test statistics, T and T^*, were obtained from LR tests with varying degrees of freedom (df) (Rospleszcz et al.; 2016).

In contrast to the studies of Rospleszcz et al. (2016), the studies in Janitza, Binder and Boulesteix (2016) assess the increase in type I error in a real data application, in which

some of the predictors might be associated with the response. For this purpose the NHANES data was used; see Appendix B for a description of the NHANES data. In addition to the setting with associations, settings with realistic data were investigated where none of the predictors was associated with the response. To obtain data sets without any associations the response of the NHANES data was randomly permuted to break any potential association between the 28 covariates and the response. This was repeated 1000 times to obtain a total of 1000 data sets in which no associations are present. The resulting data sets are called "permuted NHANES data" to distinguish them from the original NHANES data with unpermuted response.

The association between CRP level and each of the 28 covariates was univariately tested by means of an LR test. The LR test was performed to test if the full model containing the intercept and covariate $X_j, j \in \{1, 2, \ldots, 28\}$ gives a better fit than the submodel containing only the intercept. An association was considered significant if the p-value was equal to or less than 0.05. The association between the response and each variable $X_j, j = 1, \ldots, 28$, was tested in the original NHANES data sets and in bootstrap samples drawn from the original NHANES data sets. For the unpermuted NHANES data the associations were tested in $B = 10000$ bootstrap samples, and for the 1000 permuted NHANES data sets tests were performed in $1000 \times B$ bootstrap samples. Figure 5.2 shows the relative frequencies of significant associations in the bootstrap samples for the unpermuted (left panel) and the permuted (right panel) NHANES data. For bootstrap samples drawn from the unpermuted NHANES data on average (taken over $B = 10000$ bootstrap samples) there were 18.4 significant associations, while in the original unpermuted NHANES data 17 of the 28 associations were significant. For the bootstrap samples of the permuted NHANES data the average number (taken over all $1000 \times B$ bootstrap samples) of significant associations according to bootstrapped p-values was 6.12. For the original permuted NHANES data in contrast, there were on average 1.36 significant associations (over 1000 original samples).

The same computations were performed using subsamples instead of bootstrap samples, with results shown in Figure 5.3. From theory it is clear that p-values obtained from subsamples systematically deviate from p-values obtained for the original sample due to the smaller sample size and the decreased power to detect associations in subsamples: this is clearly seen in Figure 5.3. On average 14.7 of the 28 covariates were significantly associated with the CRP level in subsamples compared to 17 significant associations in the original sample.

In the case where no associations exist – the NHANES data with permuted response – a comparable number of significant findings can be observed in subsamples and in the 1000 original samples: there were on average 1.40 significant associations in subsamples

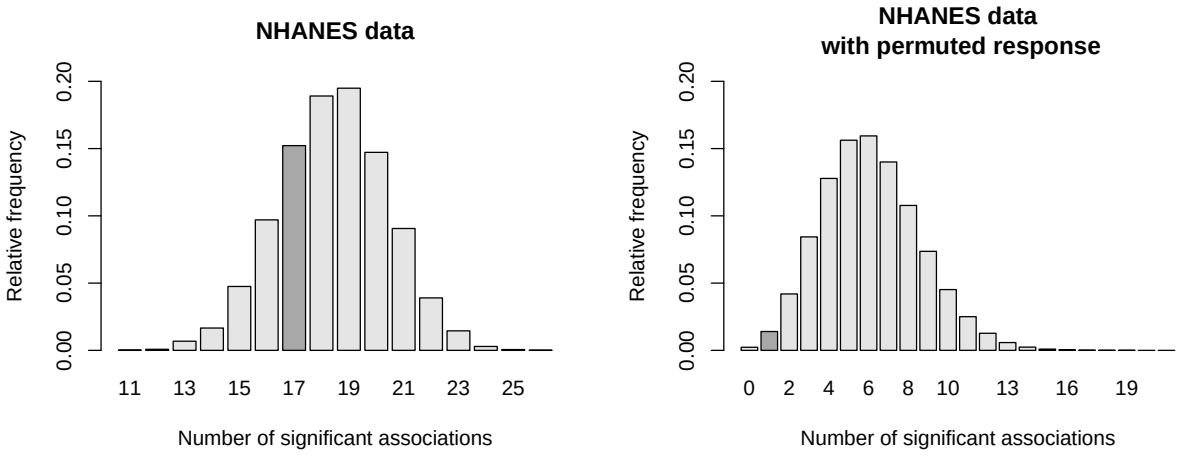

Figure 5.2.: Relative frequency of bootstrap samples with specified number of significant results when univariately testing the association between CRP level and 28 covariates. The total number of bootstrap samples was 10000 for the unpermuted NHANES data, and 10000 $\times$ 1000 for the permuted NHANES data. The dark gray bar indicates the number of significant associations in the unpermuted NHANES data (left) and the average number of significant associations in the 1000 permuted NHANES data sets (right).

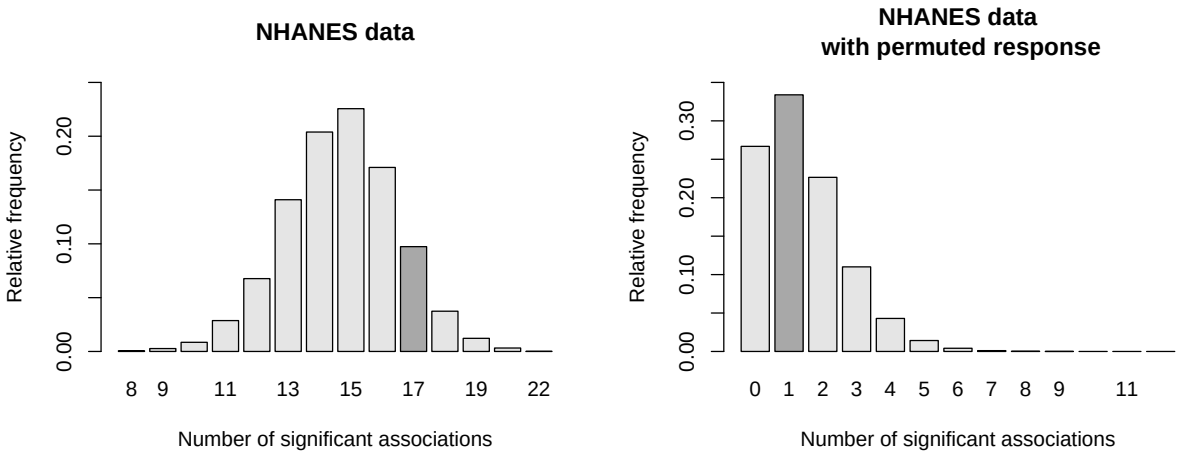

Figure 5.3.: Relative frequency of subsamples with specified number of significant results when univariately testing the association between CRP level and 28 covariates. The total number of bootstrap samples was 10000 for the unpermuted NHANES data, and 10000 $\times$ 1000 for the permuted NHANES data. The dark gray bars indicate the number of significant associations in the unpermuted NHANES data (left) and the average number of significant associations in the 1000 permuted NHANES data sets (right).

compared to 1.36 significant findings in the 1000 original samples. This is in line with the fact that tests performed on subsamples – in contrast to tests performed on bootstrap samples – do not have increased type I error, as shown for example by Sauerbrei et al. (2011). Accordingly, p-values derived on subsamples may be used for testing a specific hypothesis.

5.2.2. Distribution of bootstrapped p-values

Note that in Section 5.2.1 the marginal distribution of the bootstrapped test statistic was considered to prove that the type I error is increased when using bootstrapped p-values. However, a marginal representation does not give any information on the distribution of bootstrapped p-values for a specific sample $\boldsymbol{x}$. Moreover, it does not provide any information on whether the bootstrapped p-value can be expected to be similar to the p-value of an observed sample $\boldsymbol{x}$. These issues are addressed in the following for both the Z-test and the LR test.

Z-test

Let us consider the setting of normally distributed variables and the null hypothesis which states that the population mean equals μ_0. Now let Z be the test statistic computed based on the observed sample $\boldsymbol{x}$, and let Z^* be the bootstrapped test statistic which follows a $N(Z, 1)$ distribution conditional on $\boldsymbol{x}$ (cf. Section 5.2.1). Figure 5.4 shows distributions for $Z^*|\boldsymbol{x}$, i.e., the distributions are conditional on the sample $\boldsymbol{x}$. Conditional distributions are shown for three realizations of $\boldsymbol{x}$ with corresponding absolute Z values, namely (a) a large absolute Z value, (b) a small absolute Z value and (c) an intermediate absolute Z value. From this illustration it can be seen that the distribution of bootstrapped p-values – and with that, the discrepancy between bootstrapped p-values and the original p-value – depends on the realized sample and the respective test statistic Z:

(a) If the observed $|Z|$ is large (upper panel of Figure 5.4), there is approximately a 50% chance of having a bootstrapped p-value, p^*, which is larger than p, the observed p-value based on $\boldsymbol{x}$ (indicated by the dark gray area), and a 50% chance of having a smaller p^* (light gray area). In this scenario p^* would be considered to be a good approximation of p.

(b) If the observed value for $|Z|$ is close to 0 (middle panel), Z^* follows approximately a standard normal distribution, and the bootstrapped p-values are uniformly distributed on $[0, 1]$. Thus a p^* of 0.5 (i.e., the expectation or median of a variable $U \sim U[0, 1]$) is expected. However, p is 1. In cases where Z is close to 0, the bootstrapped p-value is obviously not a good approximation of the p-value of the original sample.

(c) If $|Z|$ takes an intermediate value, say, 1 (lower panel), the situation is similar to (a). However, in contrast to (a), there is a moderate probability for negative Z^* values smaller than $-|Z|$, or, in mathematical terms, $P(Z^* < -|Z||\boldsymbol{x})$ is much larger than 0 and cannot be ignored. Therefore, the probability of obtaining $p^* < p$ is greater

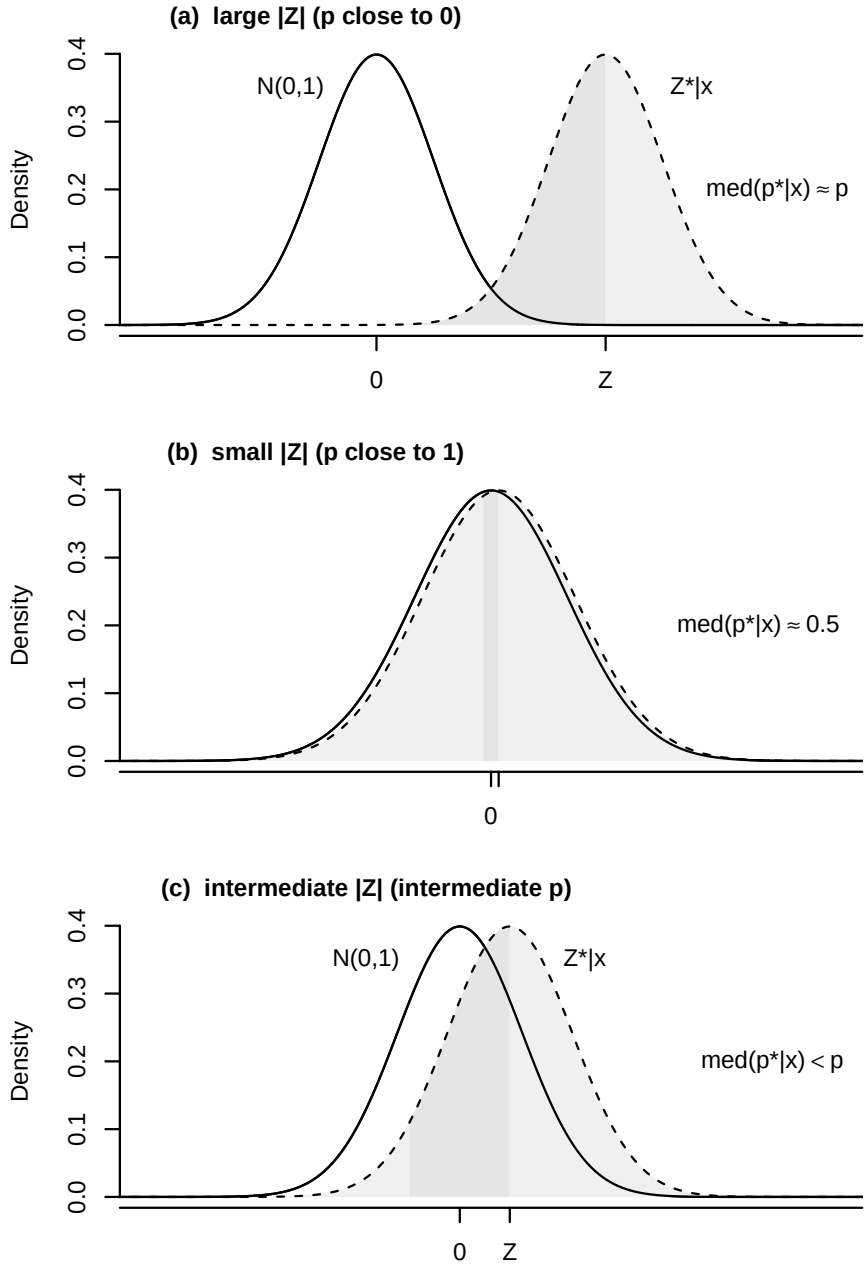

Figure 5.4.: Conditional distribution of Z^* for a fixed sample $\boldsymbol{x}$ with observed test statistic Z. Three different scenarios are considered: (a) $|Z|$ is large, (b) $|Z|$ is small, and (c) $|Z|$ is intermediate. The standard normal distribution is indicated by the solid black line. The light (dark) gray area represents the bootstrapped test statistics Z^* with corresponding bootstrapped p-value smaller (larger) than the p-value derived for the observed test statistic Z.

than obtaining $p^* > p$. This shows that bootstrapped p-values are not a good approximation of the p-value of the original sample if $|Z|$ takes an intermediate value: in over 50% of the bootstrap samples one would expect p^* smaller than p.

To summarize, the smaller the $|Z|$ (or, the larger the p), the greater the difference between the median bootstrapped p-value and p. As $|Z|$ tends to infinity (or, p tending to 0), the

difference becomes smaller. Empirical studies support these findings (not shown).

Note, however, that these considerations are for the more commonly used two-sided test, but do not generalize to the one-sided Z-test. In the following the one-sided Z-test with null hypothesis $H_0 : \mu \leq \mu_0$ and alternative hypothesis $H_1 : \mu > \mu_0$ is considered. Three hypothetical scenarios are considered: in scenario (a) a positive value for the test statistic Z is observed, in (b) a negative Z value is observed, and in (c) Z is close to 0. The respective conditional distributions of bootstrapped test statistics are shown in Figure 5.5. The light (dark) gray area represents the bootstrapped test statistics Z^* with corresponding bootstrapped p-value smaller (larger) than the p-value derived for the observed test statistic Z. For all scenarios exactly 50% of the bootstrapped p-values are expected to be larger and 50% smaller than the p-value based on the original sample (i.e., the median bootstrapped p-value is close to the p-value computed from Z). To conclude, for the one-sided Z-test bootstrapped p-values give a good approximation of the p-values computed from the original data.

One might argue that it was already shown in Section 5.2.1 that bootstrapped p-values do not approximate the original p-values very well. It is important to note that the increased type I error does not imply that bootstrapped p-values are a poor approximation of original p-values. For the one-sided Z-test for example, tests performed using bootstrapped p-values have increased type I error, but bootstrapped p-values are a good approximation of the originals.

Likelihood Ratio test

The considerations made for the two-sided Z-test apply to the LR test in a similar way. Let us assume for the moment that the null hypothesis (that the submodel is true) holds and that the LR test statistic T equals zero for an observed sample, which means that in this observed sample the derived likelihood of the submodel is exactly equal to the likelihood of the unrestricted model. Then the bootstrap samples are drawn from a distribution in which H_0 is true. Accordingly, the bootstrapped test statistic T^* follows a central χ^2-distribution and the bootstrapped p-value is uniformly distributed on $[0, 1]$. As with the two-sided Z-test, the median and expectation of the bootstrapped p-value is 0.5, while the p-value for the original sample is 1. Bootstrapped p-values for the LR test thus cannot be expected to be close to p-values computed on the original data.

It is difficult to explore the distribution of bootstrapped p-values dependent on different values of the LR test statistic T by theoretical arguments, as it was done for the Z-test, since the conditional distribution of T^* given the observed sample is unknown. Therefore, the discrepancy between bootstrapped p-values and original p-values was further investigated using empirical studies of the NHANES data. LR tests for each of the 28

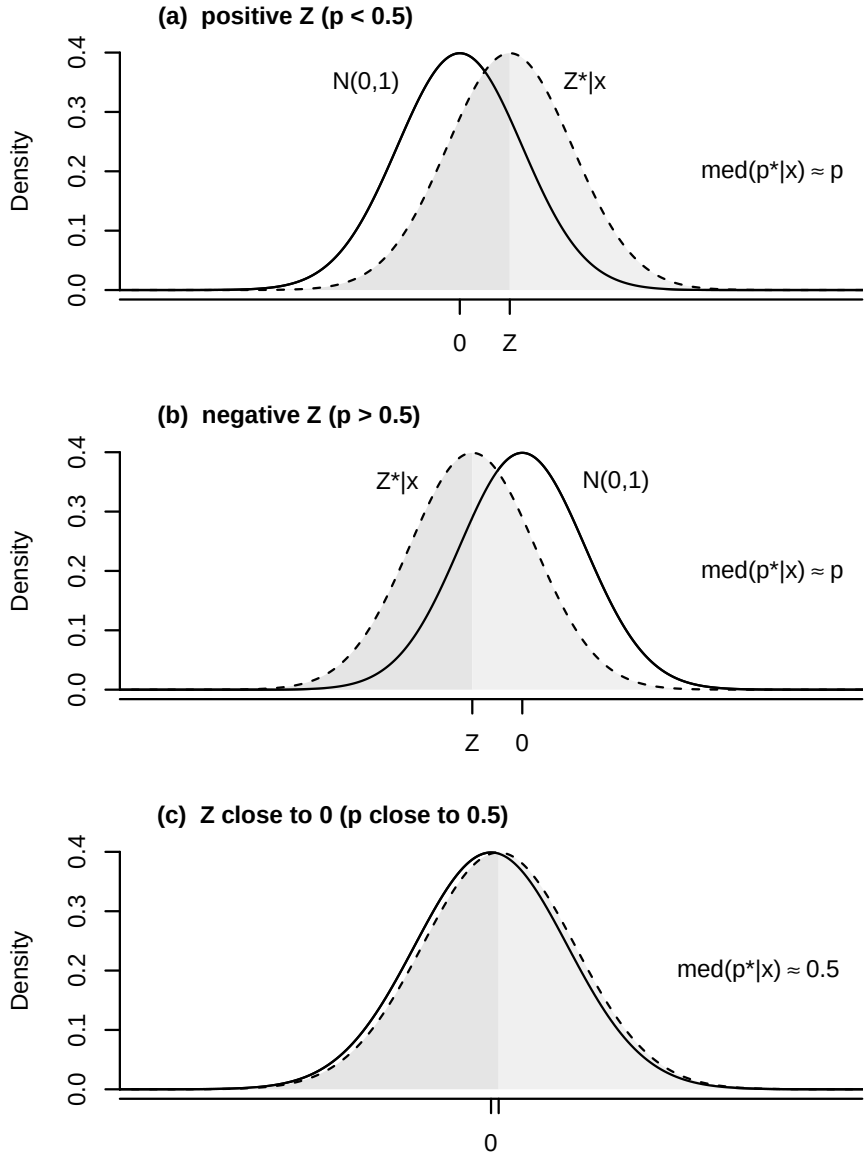

Figure 5.5.: Conditional distribution of Z^* for a fixed sample $\boldsymbol{x}$ with observed test statistic Z when performing a one-sided test with hypotheses $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$. Three scenarios are considered: (a) Z is positive, (b) Z is negative, and (c) Z is close to 0. The standard normal distribution is indicated by the gray solid line. The light (dark) gray area represents the bootstrapped test statistics Z^* with corresponding bootstrapped p-value smaller (larger) than the p-value derived for the observed test statistic Z.

variables were performed to test the null hypothesis (the submodel containing only the intercept is true) against the alternative hypothesis (the model containing the intercept plus the respective variable is true). This was done for the original data as well as for 10000 bootstrap samples and 10000 subsamples drawn from each original data set. As original data sets both the unpermuted NHANES data and 1000 permuted NHANES data sets were used. Figure 5.6 (left panel) shows the median bootstrapped p-values for each of the 28 variables plotted against the p-values obtained for the original sample.

Black points represent the p-values for LR tests with 1 degree of freedom (performed for metric and binary variables), while the gray points correspond to tests with 3 or more degrees of freedom (for categorical variables with 4, 5, 6 or 12 categories). For the sake of clarity the results are shown only for the first 10 permuted NHANES data sets (right panel). Note that since there are 10 data sets, 10×28 points are plotted. For LR tests with 1 degree of freedom a similar situation to that for the Z-test is observed: when the p-value is small, it is approximated well by the median bootstrapped p-value; however, for large p-values the approximation is not good. For LR tests with 3 or more degrees of freedom it seems bootstrapped p-values are never a good approximation, independent of whether the original p-values are small or large.

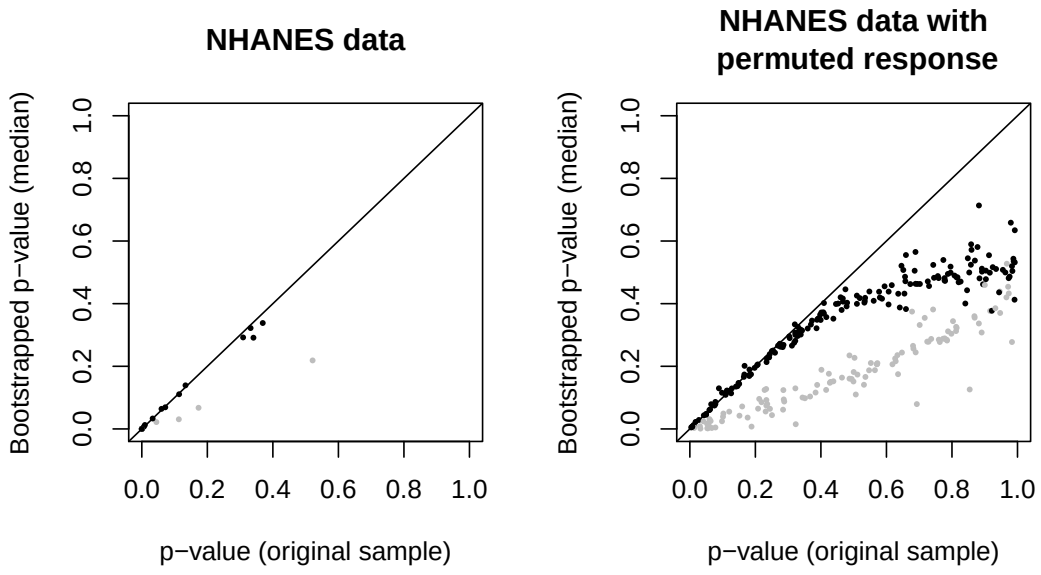

Figure 5.6.: Median p-values obtained for testing the association between CRP level and each of the 28 covariates in 10000 bootstrap samples, plotted against the p-values of the original sample, for the NHANES data. Black points represent the p-values for LR tests with 1 degree of freedom, and gray points correspond to LR tests with 3 or more degrees of freedom. Points lying on the diagonal line would indicate agreement between p-values derived on the original NHANES data and bootstrapped p-values. Left: Results obtained for the unpermuted NHANES data. Right: Results obtained for 10 permuted NHANES data sets in which there are no true associations between covariates and the CRP level (via permuting values for CRP level).

Figure 5.7 shows the median subsampled p-values plotted against the p-values obtained for the original sample. It can be observed that subsampled p-values were larger than p-values for the original sample if the latter were in the range $[0, 0.6]$. If p-values obtained for the original sample were above 0.6, subsampled p-values were smaller. It is clear that due to a different sample size subsampled p-values are not a good approximation of p-values for original samples.

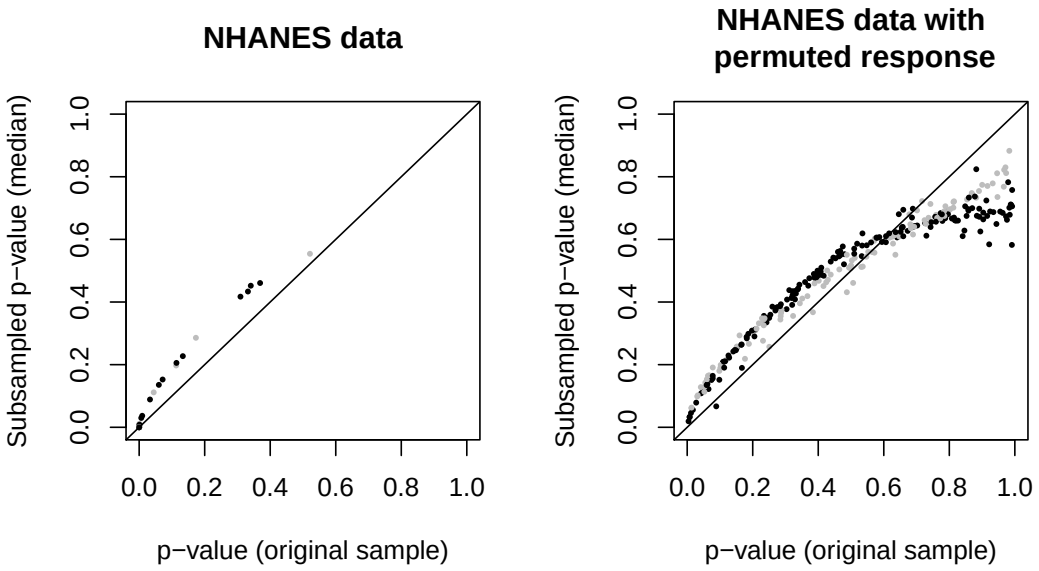

Figure 5.7.: Median p-values obtained for testing the association between CRP level and each of the 28 covariates in 10000 subsamples, plotted against the p-value of the original sample, for the NHANES data. Black points represent the p-values for LR tests with 1 degree of freedom, and gray points correspond to LR tests with 3 or more degrees of freedom. Points lying on the diagonal line would indicate agreement between p-values derived on the original NHANES data and p-values derived on subsamples. Left: Results obtained for the unpermuted NHANES data. Right: Results obtained for 10 permuted NHANES data sets in which there are no true associations between covariates and the CRP level (via permuting values for CRP level).

5.3. Application 1: Bootstrapped p-values for multivariable model building

Multivariable regression is commonly used in biometrical applications to model the association between an outcome and candidate predictor variables. Automated selection procedures such as stepwise selection, forward selection or backward elimination, are often used in this context. These methods are usually based on information criteria such as the Akaike Information Criterion (AIC; Akaike; 1973), or on hypothesis tests such as the LR test. However, such model selection strategies are known to be highly unstable since small changes in the data might lead to a substantially different model. A popular bootstrap-based method which is often applied in biometrical applications investigates the stability of stepwise model selection procedures. This procedure makes use of the bootstrap to generate pseudo-samples, and model selection is performed on each bootstrap sample, where p-values of the LR test are used to decide on the inclusion of variables in the model (Chen and George; 1985; Altman and Andersen; 1989; Sauerbrei and Schumacher; 1992). The resulting "bootstrap models" can then be examined. Moreover, the importance of variables might be assessed from their occurrence in the models (see,

e.g., Sauerbrei and Schumacher; 1992). The proportion of bootstrap samples for which a variable is selected is often referred to as the variable's *inclusion frequency*. In addition to their use in assessing the importance of variables, the variables' inclusion frequencies also provide useful information on the model stability (see, e.g., De Bin et al.; 2016).

Rospleszcz et al. (2016) investigated models that were obtained when applying backward elimination based on the LR test on bootstrap samples. Their data settings included both metric and categorical predictor variables with different numbers of categories. In the backward elimination procedure a categorical predictor variable was eliminated from a model as a whole. This means that all parameters related to a categorical variable are either in a model or not in a model. The decision on the elimination of categorical variables from a model was made based on LR tests that test if the model including all the parameters related to a variable gives a better fit than the submodel which does not contain any of the parameters related to the variable. This global testing approach has the advantage of avoiding the possibility that categorical predictor variables with more categories are included more often in a model due to multiple testing.

Some of the results which were presented in Rospleszcz et al. (2016) are described in the following. The NHANES data was used for their investigations. Note that in contrast to the studies presented in Section 5.2, only the original, that is, unpermuted version of the NHANES data was considered. Models were obtained for the original NHANES data and for 5000 bootstrap samples and 5000 subsamples of the NHANES data. The full model from which variables were successively excluded contained all 28 variables which are potentially associated with the CRP level (see Appendix B for details on the NHANES data). For the original NHANES data a model was selected which included the 11 predictor variables *WBCcount*, *waistcircum*, *Cholesterol*, *age*, *BMI*, *alcohol*, *sex*, *AcuteIllness*, *race*, *HealthStatus* and *ToothCond* (see Table B.1 for a description of the variables).

The percentages of bootstrap and subsample models that include a specific number of predictor variables are shown in Figure 5.8. On average more predictor variables are included in models when using the bootstrap than with subsampling. Often more than 11 variables (which is the number of variables in the model for the original NHANES data) are included if models were derived from bootstrap samples. Thus models are selected on bootstrap samples which are systematically more complex (in terms of included variables) compared to models based on the original data. For subsampling in contrast, fewer variables are included in the models than in the model for the original sample. Accordingly, models derived from subsamples are less complex than models based on the original data. This is due to the smaller statistical power with subsampling. The complexity of models fit on subsamples thus does not represent the complexity of models fit on original samples, either.

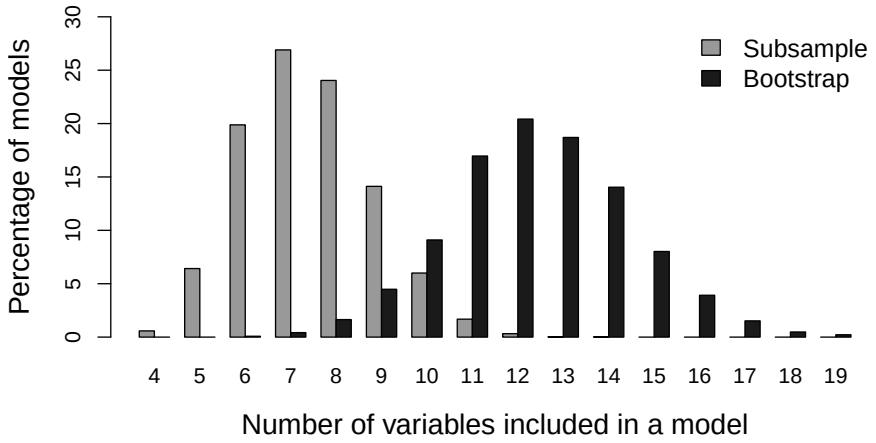

Figure 5.8.: Percentage of models with specified number of included predictor variables in subsamples and bootstrap samples (Rospleszcz et al.; 2016).

Further, the type of variables included in models derived from bootstrap samples and subsamples was compared. The simulation studies of Rospleszcz et al. (2016) showed that inclusion frequencies obtained by subsamples reliably reflect the relative importance of variables. Thus the inclusion frequencies obtained from subsamples might be considered as "gold standard" and were compared to inclusion frequencies obtained from bootstrap samples. The inclusion frequencies obtained from bootstrap samples and subsamples are shown in Table 5.3. Due to the larger number of variables in a model based on bootstrap samples, it is clear that, overall, the inclusion frequencies are higher if computed based on bootstrap samples. However, much more interesting is the finding that there are a few cases in which a binary variable has higher inclusion frequencies than a categorical variable with more than two categories for subsamples but the categorical variable with more than two categories is more frequently included than the binary for the bootstrap samples. Although the binary variable is possibly more important (since subsamples reliably reflect the relative importance of variables), the bootstrap suggests that the categorical variable with more than two categories is more important than the binary. This is due to the fact that on bootstrap samples the increase in type I error is more extreme for LR tests with many degrees of freedom (e.g., when testing a categorical variable with many categories; see Section 5.2.1). Therefore categorical variables with many categories are preferentially included in models derived from bootstrap samples. Consider as an example the binary variable *stroke* whose inclusion frequency is 17.86% for subsamples. Inclusion frequencies for the 4-category variable *depression* and the 5-category variable *sleepTrouble* are 7.60% and 12.46%, respectively, and are thus lower. Since inclusion frequencies obtained by subsampling have been shown to reliably reflect the relative importance of variables, it is assumed that the variable *stroke* is more strongly associated with CRP than the variables *depression* and *sleepTrouble*. For bootstrap

Variable	Scale	df	Inclusion frequency (in %)	
			Subsample	Bootstrap
age	metric	1	30.12	51.08
alcohol	metric	1	0.00	4.88
BMI	metric	1	99.98	99.90
BPdias	metric	1	7.46	22.36
BPsys	metric	1	20.12	45.38
Cholesterol	metric	1	4.46	21.56
waistcircum	metric	1	6.98	17.06
WBCcount	metric	1	100.00	99.96
100cig	categorical with 2 categories	1	12.36	28.68
AcuteIllness	2 categories	1	82.58	87.10
asthma	2 categories	1	6.46	17.94
chronicBronchitis	2 categories	1	0.68	11.96
diabetes	2 categories	1	9.24	35.60
heartFailure	2 categories	1	0.08	4.02
heavyDrinker	2 categories	1	0.48	5.84
sex	2 categories	1	72.54	71.12
stroke	2 categories	1	17.86	30.86
country_of_birth	4 categories	3	0.48	16.70
depression	4 categories	3	7.60	45.98
education	5 categories	4	2.28	31.26
HealthStatus	5 categories	4	41.56	69.50
medicalPlaceToGo	5 categories	4	0.02	7.56
race	5 categories	4	39.62	73.78
sleepTrouble	5 categories	4	12.46	41.20
ToothCond	5 categories	4	32.86	68.72
wakeUp	5 categories	4	38.26	68.38
marital_status	6 categories	5	51.98	74.90
income	12 categories	11	46.00	87.52

Table 5.3.: Inclusion frequencies for the predictor variables in the NHANES data for 5000 subsamples and 5000 bootstrap samples. The scale of the variables and the degrees of freedom (df) of the corresponding LR test are also shown.

samples, however, the inclusion frequency for *stroke* is 30.86% (which is higher than that for the subsamples, as one would expect) but the inclusion frequencies for *depression* and *sleepTrouble* are even higher, with 45.98% and 41.2%, respectively. In other words, if the importances of these variables were to be assessed based on bootstrap samples, the association between CRP level and the predictor variables *depression* and *sleepTrouble* would be incorrectly estimated to be higher than the association between CRP level and the predictor variable *stroke*.

There is a number of other examples in which a binary variable yields higher inclusion frequencies than a categorical variable with $m > 2$ categories for subsamples but the categorical variable with $m > 2$ is more frequently included than the binary for the bootstrap samples. All cases are specified in the following:

- *100cig* ($m = 2$) and *depression* ($m = 4$),
  and *education* ($m = 5$),
- *sex* ($m = 2$) and *race* ($m = 5$),
  and *marital_status* ($m = 6$),
  and *income* ($m = 12$),
- *stroke* ($m = 2$) and *depression* ($m = 4$),
  and *education* ($m = 5$),
  and *sleepTrouble* ($m = 5$),
- *AcuteIllness* ($m = 2$) and *income* ($m = 12$),
- *asthma* ($m = 2$) and *education* ($m = 5$),
- *diabetes* ($m = 2$) and *depression* ($m = 4$).

In contrast to that, there is no case where a categorical variable with more than two categories has a higher inclusion frequency than a binary or metric variable for the subsamples, but is less frequently included than the binary or metric variable on bootstrap samples. Overall, these findings show that depending on the number of categories the effect of categorical variables might be considerably overestimated when applying automated selection procedures on bootstrap samples.

5.4. Application 2: Bootstrapped p-values for variable ranking

This section investigates the consequences when using bootstrapped p-values for ranking variables. Such an approach has been proposed by Mukherjee et al. (2003) for ranking genes with respect to their differential expression. For each variable, bootstrapped p-values are computed for testing the null hypothesis that the variable is not associated with the response. A ranking is obtained by sorting the variables by their mean or median bootstrapped p-value. Such resampling-based strategies might be helpful in minimizing the influence of some extreme observations which may greatly affect a variable ranking. Mukherjee et al. (2003) proposed their variable ranking approach in the context of ranking genes in small samples, where influential points are likely to have a huge impact. In the following this approach is applied to the NHANES data to obtain a ranking of the 28 considered variables. For each variable a LR test was performed to test the null hypothesis that the variable is not associated with the CRP level. Both the unpermuted NHANES data and 1000 permuted NHANES data sets were considered. The ranks were made based on the median bootstrapped p-value. Figure 5.9 shows the variable rankings for the unpermuted NHANES data. The upper left panel corresponds to the rankings

according to the p-values from the original NHANES sample and the upper right panel corresponds to rankings by the median bootstrapped p-values (i.e., the median of 10000 bootstrapped p-values). In addition, results are shown when using the median p-value obtained from 10000 subsamples (lower panel). On the whole, the rankings are similar, especially among those variables with strong evidence for association. However, close inspection reveals some differences between the rankings based on the original sample and those based on bootstrap samples. These differences are likely attributable to the fact that categorical variables with many categories obtain systematically smaller bootstrapped p-values than metric variables or categorical variables with fewer categories. This leads to a ranking in which variables with many categories gain ranking positions closer to the top when ranked by the median bootstrapped p-value. Table 5.4 shows the ranking positions for each variable separately. There are many cases in which categorical variables with four or more categories gain ranking positions closer to the top when ranked by bootstrapped p-values. Conversely, the binary and metric variables are located at positions at the bottom of the ranking when the ranking is according to bootstrapped p-values. In contrast, when using subsamples there are only minor differences in the ranking, with seemingly no effect of a variable's scale on its ranking position.

The observed mechanisms are even more extreme for the permuted NHANES data sets; see Table C.1, which shows the result for the first permuted data set. For the permuted data sets there are very large differences in the variable ranking – with variables with many categories ranked at top positions and binary or metric variables at much lower positions when p-values are derived from bootstrap samples.

To conclude, the studies show that, though resampling procedures might be promising methods for obtaining stable variable ranking lists, bootstrapped p-values should not be compared with significance thresholds for making decisions on the significance of variables. In particular, care needs to be taken when the interest lies in ranking variables of different scales, which often occurs in epidemiological studies. An example of further relevance is gene ranking when single nucleotide polymorphisms are considered, which for some genes are represented by a categorical variable with three categories but for others only two categories. Moreover, associations between genes and a phenotype are usually weak or non-existent, which is expected to be especially problematic as suggested by the results of the permuted NHANES data. Thus, bootstrapped p-values should not be applied for obtaining variable rankings in settings including categorical variables. Subsampling may be a reasonable alternative to the bootstrap for variable ranking: In the studies there were only minor differences between the rankings that were obtained by sorting variables by p-values obtained from the original sample and from subsamples. This might indicate that in the considered NHANES data there are not many influential

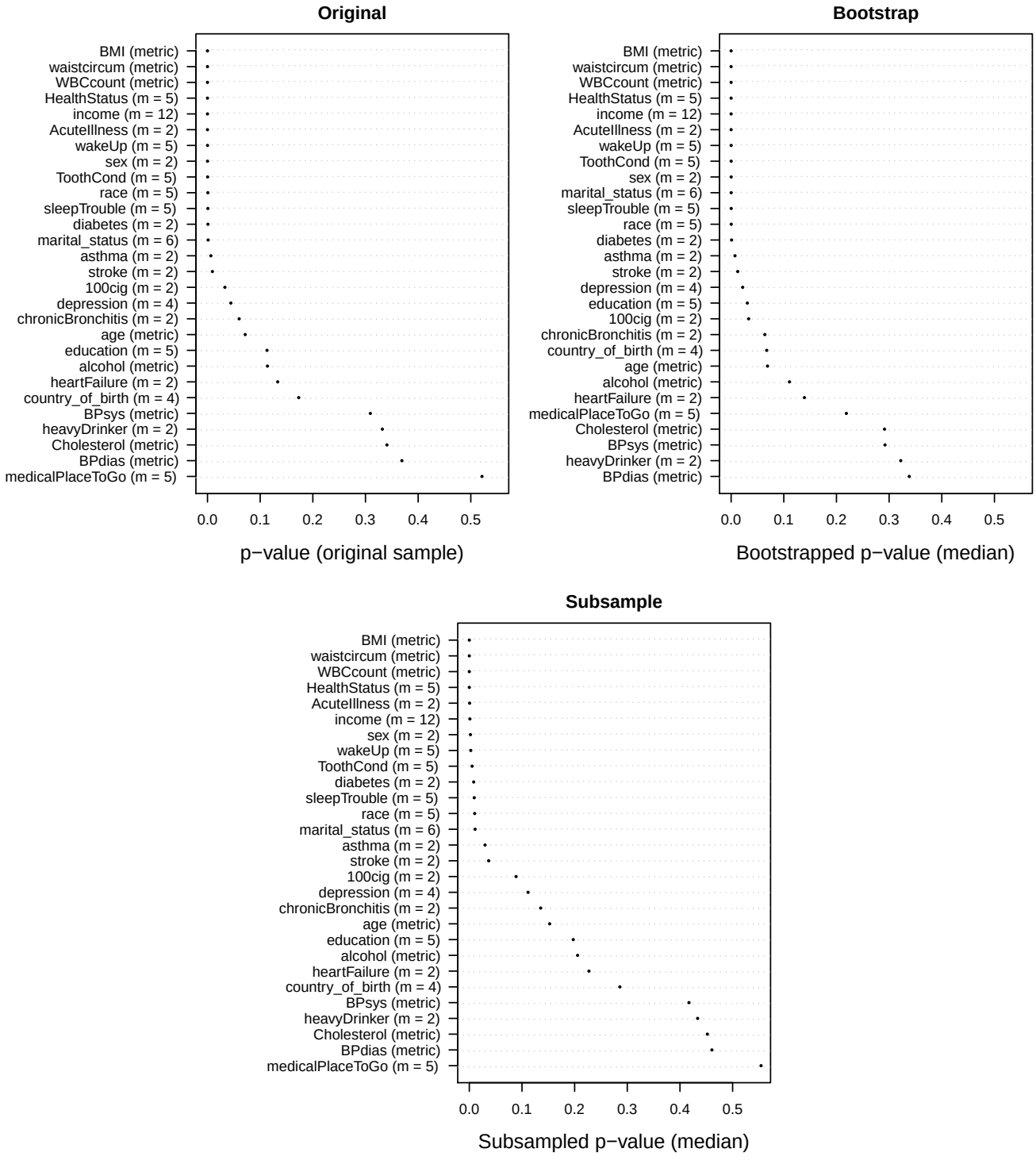

Figure 5.9.: Variable ranking by p-values obtained for the original unpermuted NHANES sample (upper left) and by the p-value obtained from the median over 10000 bootstrapped p-values (upper right) and the median p-value from subsamples (lower). The parameter m denotes the number of categories of a categorical variable.

points that have a large impact on the results, but more research is needed on this topic. To conclude, subsampling should be preferred over bootstrapping for obtaining variable rankings if variables are of different scales. However, in settings with very small sam-

Scale	Variable	Original rank	Bootstrap rank	(diff.)	Subsample rank	(diff.)
metric or $m = 2$	BMI	1	1	(0)	1	(0)
	waistcircum	2	2	(0)	2	(0)
	WBCcount	3	3	(0)	3	(0)
	AcuteIllness	6	6	(0)	5	(+1)
	sex	8	9	(−1)	7	(+1)
	diabetes	12	13	(−1)	10	(+2)
	asthma	14	14	(0)	14	(0)
	stroke	15	15	(0)	15	(0)
	100cig	16	18	(−2)	16	(0)
	chronicBronchitis	18	19	(−1)	18	(0)
	age	19	21	(−2)	19	(0)
	alcohol	21	22	(−1)	21	(0)
	heartFailure	22	23	(−1)	22	(0)
	BPsys	24	26	(−2)	24	(0)
	heavyDrinker	25	27	(−2)	25	(0)
	Cholesterol	26	25	(+1)	26	(0)
	BPdias	27	28	(−1)	27	(0)
$m = 4$	depression	17	16	(+1)	17	(0)
	country_of_birth	23	20	(+3)	23	(0)
$m = 5$	HealthStatus	4	4	(0)	4	(0)
	wakeUp	7	7	(0)	8	(−1)
	ToothCond	9	8	(+1)	9	(0)
	race	10	12	(−2)	12	(−2)
	sleepTrouble	11	11	(0)	11	(0)
	education	20	17	(+3)	20	(0)
	medicalPlaceToGo	28	24	(+4)	28	(0)
$m = 6$	marital_status	13	10	(+3)	13	(0)
$m = 12$	income	5	5	(0)	6	(−1)

Table 5.4.: Variable ranking for the unpermuted NHANES data. Variable rankings are obtained from p-values obtained for the original NHANES sample ("Original rank"), from the median bootstrapped p-value ("Bootstrap rank"), and from the median p-value from subsamples ("Subsample rank"). The difference to the "Original rank" is given in brackets for each variable. The parameter m denotes the number of categories of a categorical variable.

ple sizes – for which the ranking approach was originally proposed (Mukherjee et al.; 2003) – subsampling from a data set that consists of only a few observations may not be advisable.

5.5. Application 3: Bootstrapped p-values for assessing the variability of p-values

In their paper, Boos and Stefanski (2011) propose reporting the variability of p-values conjointly with the p-value in real data applications to gain a better understanding of the variability of p-values if one were to replicate this study. They propose approximating the variance of p-values, or preferably the variance of $-\log_{10}(p\text{-value})$, based on bootstrapped p-values. If the standard deviation is a large fraction of the p-value, there is

high variability, which may explain the fact that identical experiments may lead to rather distinct p-values. The question arises of whether the variance of bootstrapped p-values can be used to approximate the variability of p-values that would be observed if the same experiment was repeatedly performed. To investigate this issue simulation studies were performed, which allow drawing multiple times from the true distribution.

Simulation studies were performed for both the Z-test and the LR test, and two settings were considered: a setting, in which the null hypothesis is true and a setting, in which the alternative hypothesis is true. The simulation studies are outlined in the following:

Z-test: In the setting, in which the null hypothesis is true, $x_i, i = 1, \ldots, 1000$ were independently drawn from $N(0,1)$. The null hypothesis $H_0 : \mu = 0$ was tested against the alternative hypothesis $H_1 : \mu \neq 0$. A total of 10000 bootstrap samples were generated by drawing from this sample with replacement. A Z-test was performed for the original sample and for each bootstrap sample. Subsequently the standard deviation of the 10000 bootstrapped p-values and the standard deviation of the negative logarithm of the bootstrapped p-values were computed. This process was repeated 10000 times and the standard deviation of the p-values and that of the negative logarithm of the p-values for the 10000 original data sets were computed. The same analysis was performed for a setting where the alternative hypothesis is true; in this setting $x_i, i = 1, \ldots, 1000$ were independently drawn from $N(0.08, 1)$.

LR test: In this study $x_{i1}, \ldots, x_{i,10}$ were independently drawn for $i = 1, \ldots, 1000$ from a multivariate normal distribution with expectation $\boldsymbol{\mu} = (0, \ldots, 0)^\top \in \mathbb{R}^{10}$ and covariance matrix $\boldsymbol{I}_{10}$, corresponding to the identity matrix of dimension 10×10. The response variable Y_i was generated according to the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{10} x_{i,10} + \epsilon_i,$$

with $\epsilon_i \sim N(0,1)$ for $i = 1, \ldots, 1000$. The global null hypothesis is that none of the predictor variables is associated with the response, that is $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{10} = 0$. The alternative hypothesis is that at least one of them is associated, that is $H_1 : \beta_j \neq 0$ for at least one $j \in \{1, \ldots, 10\}$. The corresponding LR test compares the likelihood of the submodel L_0 which contains only the intercept, to the likelihood L_1 of the model which includes all predictor variables $X_1, \ldots, X_{10}$. If the null hypothesis is true the LR test statistic (5.6) follows a central χ^2-distribution with 10 degrees of freedom. In the first setting the null hypothesis is true, that is $\beta_j = 0$ for all $j = 1, \ldots, 10$. In the second setting, the alternative hypothesis is true and the coefficients were $\beta_j = 0.02$ for $j = 1, \ldots, 10$.

The derivation of the standard deviations of p-values and standard deviations of $-\log_{10}(p\text{-value})$ based on original samples and bootstrap samples was exactly the same as described for the Z-test.

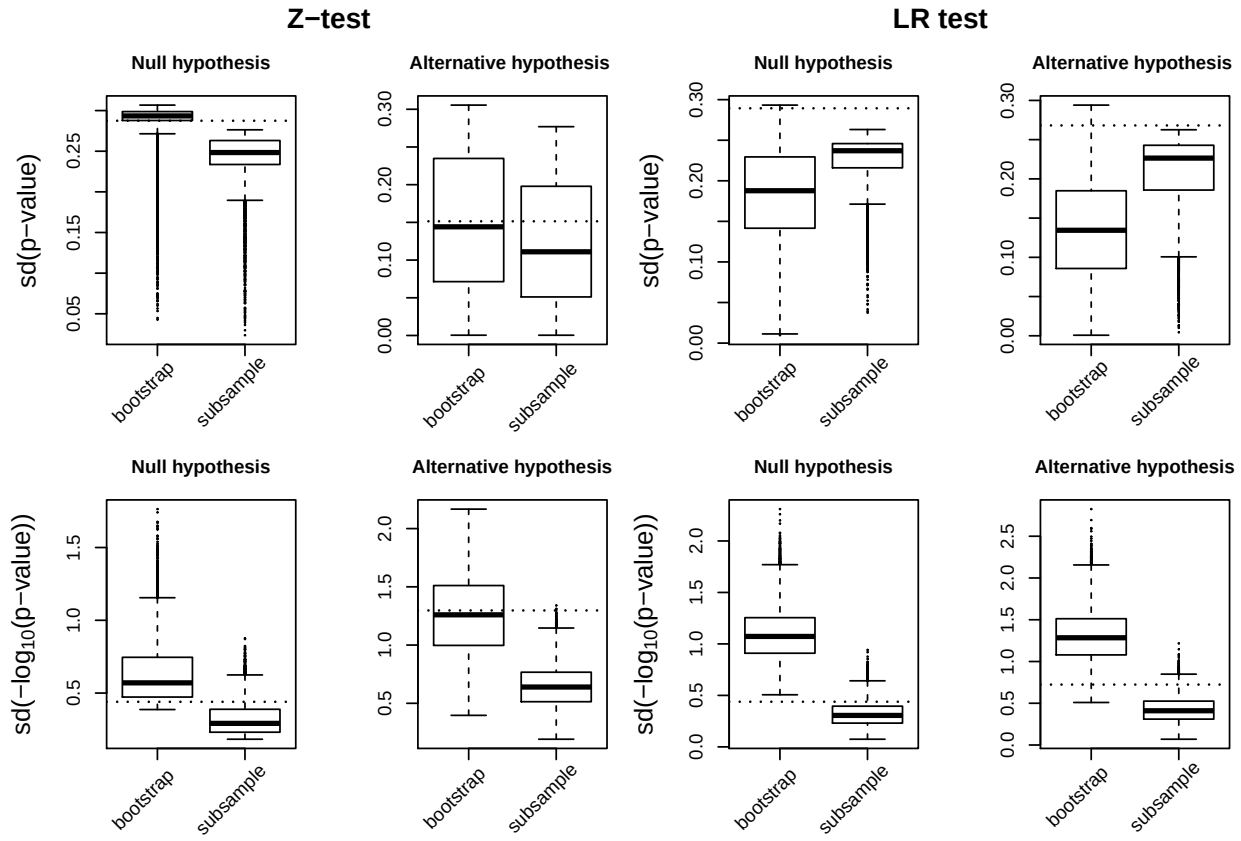

Figure 5.10.: Standard deviations (sd) of bootstrapped and subsampled p-values or $-\log_{10}(p$-value) for the Z-test (left two columns) and the LR test with 10 degrees of freedom (right two columns). The dotted line represents the standard deviation of the p-values or negative logarithm of the p-values, respectively, computed from the 10000 original samples. Each original sample gave rise to 10000 bootstrap samples and 10000 subsamples.

Figure 5.10 shows the distributions of the standard deviations of the bootstrapped p-values and the standard deviations of the negative logarithm of the bootstrapped p-values for the Z-test and the LR test. The dotted line represents the standard deviation of the p-values or negative logarithm of the p-values computed from the 10000 original samples. Results are also shown for subsampling. For the Z-test under H_0 a systematic but probably negligible difference is observed between the standard deviations of p-values for the original data and those of bootstrapped p-values (or $-\log_{10}(p\text{-value})$). Under H_1, in contrast, the standard deviation of bootstrapped p-values (or $-\log_{10}(p\text{-value})$) seems to be a good approximation of the true p-value variability. For the LR test with 10 degrees of freedom the standard deviations of the p-value (or $-\log_{10}(p\text{-value})$) computed on bootstrap samples do not reflect the true p-value variability in the studies, neither under the null hypothesis nor under the alternative hypothesis. This result was expected

in the light of the empirical results presented in Section 5.2.2, which showed that for LR tests with 3 or more degrees of freedom the distribution of bootstrapped p-values is always different from the true p-value distribution. Subsampling is not a reasonable alternative here since tests performed on subsamples do not reflect the p-value variability of the original samples either, as seen in Figure 5.10.

5.6. Discussion

In this chapter, it was shown through theoretical and empirical results that when using bootstrapped p-values there is increased type I error for both the Z-test and the LR test, and that the increase in type I error also depends on the degrees of freedom of the LR test. Similar results are expected for other statistical tests. Investigations on the conditional distribution of bootstrapped p-values showed that for the one-sided Z-test bootstrapped p-values give a good approximation of the p-values computed from the original data, but that this is generally not the case for the two-sided Z-test and the LR test. This shows that despite the fact that the type I error is increased when using bootstrapped p-values, bootstrapped p-values might be a good approximation of the originals.

Three practices making use of bootstrapped p-values were investigated: bootstrapping p-values for multivariable model building, for variable ranking and for assessing the variability of p-values. The three approaches were applied on simulated data or on a large real data set from a population-based study, respectively, to investigate whether results are affected by the problems mentioned above. When backward elimination based on the LR tests was performed using bootstrap samples, the resulting models included more parameters than the model which was obtained for the original sample. This is likely due to the increased type I error in bootstrap samples. However, other characteristics of the bootstrap may also play an important role, such as the replication of possible influential points or outliers in bootstrap samples. Further research is needed to address this issue.

Moreover, the studies presented in this chapter showed that bootstrapped p-values should not be used for multivariable model building if there are variables of different scales because categorical variables with more categories are favored over variables with fewer categories and over metric variables. This is related to the fact that on bootstrap samples the increase in type I error is more extreme for LR tests with many degrees of freedom. The same problem applies when ranking variables by bootstrapped p-values obtained from LR tests: In settings without any associations, variables with many categories were ranked at top positions and binary or metric variables at much lower positions when p-values were derived from bootstrap samples. The studies suggest that this problem is especially pronounced in settings with weak associations, and is less pro-

nounced in settings with very strong associations. Finally, the variability of bootstrapped p-values was shown not to reflect the variability of p-values for the LR test when repeating the same experiment several times, thus making the reliability of the approach suggested by Boos and Stefanski (2011) questionable.

The use of subsampling was investigated as a promising alternative strategy to circumvent problems induced by the bootstrap. The properties of subsampling have been theoretically investigated in the literature; it has been shown that subsampling has desirable properties even in situations where the bootstrap fails. A recent approach to stability selection based on subsampling was introduced by Meinshausen and Bühlmann (2010). Their studies impressively show that subsampling is a powerful tool in investigating the stability of models, such as penalized likelihood models and graphical models. Further, Strobl et al. (2007) proposed the use of subsampling instead of bootstrapping in the context of random forests to circumvent the problem of preferential selection of certain types of predictors for a split, and Rospleszcz et al. (2016) showed that the inclusion frequencies obtained from subsamples reliably reflect the relative importance of variables and should thus be preferred over bootstrap inclusion frequencies.

However, the results in this chapter show that subsampling should not be regarded as an universally applicable alternative to the bootstrap. For investigating the variability of p-values, for example, subsampling is not appropriate. These investigations make it clear that subsampling is not a reliable alternative to the bootstrap for all types of applications, even if it has shown important advantages in some situations (Strobl et al.; 2007; Rospleszcz et al.; 2016; De Bin et al.; 2016).

6. Model selection through information criteria and data splitting approaches on bootstrap samples

Fitting a prediction model and evaluating its prediction error on the same data is not a trivial task, especially if the model involves one or several tuning parameters. To avoid overoptimism a data splitting procedure should be applied in which the model is fit on one part of the data and evaluated on the other part of the data (see, e.g., Boulesteix et al.; 2008). One option is to use a bootstrap sample to fit the model (*model building step*) and to use the remaining observations which were not part of the bootstrap sample (out-of-bag observations) to compute the model's prediction error (*model evaluation step*). This process is usually repeated a large number of times and the average error over the replications is obtained. If the statistical model integrates tuning parameters such as the number of boosting steps for gradient boosting algorithms (Friedman; 2001; Bühlmann and Hothorn; 2007), the optimal value for a tuning parameter is often determined by using information criteria or through application of an internal cross-validation procedure. However, it was shown that information criteria derived from bootstrap samples systematically deviate from information criteria derived from original samples (Wagenmakers et al.; 2004; Steck and Jaakkola; 2003). But the practical consequences of this systematic deviation are largely unknown. It is also unknown if cross-validation procedures applied on bootstrap samples are useful for selecting appropriate values for tuning parameters. Practical consequences can relate to several different aspects. For example, they may relate to structural differences in the models obtained from bootstrap samples and models obtained from original samples, where the differences are due to the selection of different optimal values for the tuning parameter. Differences in the models' structure may lead to wrong conclusions regarding, for example, the complexity of the considered relationship, the effect of single predictor variables or the predictive value of the combination of variables included in a model. This chapter addresses the practical consequences with focus on model complexity and model accuracy. Section 6.1 deals with tuning parameter selection through information criteria. This part is based on Janitza, Binder and Boulesteix

(2016) but also shows simulation studies not presented therein. Smaller parts of the simulation studies were shown in the former version of the article which is available as a technical report (Janitza et al.; 2014). In Section 6.2 the use of cross-validation procedures for tuning parameter selection is addressed. Some results of the simulation studies of this section were described in Janitza et al. (2014). In both sections gradient boosting algorithms are considered as an example. Promising alternatives to the bootstrap, such as subsampling, are also considered.

6.1. Tuning parameter selection through information criteria

Information criteria are often used for the comparison of non-nested models. These measures compare models based on their goodness-of-fit to the data while penalizing the complexity of the model (see also Burnham and Anderson; 2002). Akaike's information criterion (AIC) is a widely used measure for model selection. It is defined as

$$\text{AIC} = -2\log(L) + 2p, \tag{6.1}$$

where L denotes the likelihood and p denotes the number of parameters included in the model. It has been shown that minimizing the AIC is approximately equivalent to minimizing the expected Kullback-Leibler distance between the true and the estimated density (Akaike; 1973).

The bootstrapped AIC is given by

$$\text{AIC}^* = -2\log(L^*) + 2p, \tag{6.2}$$

with L^* denoting the likelihood computed for a model that was fit on a bootstrap sample. To prove that the bootstrapped AIC is not a good approximation of the AIC defined in (6.1), two nested models differing in the inclusion of only one parameter will be compared [1]. If AIC_1 denotes the AIC of the unrestricted model that includes p parameters and AIC_0 denotes the AIC of the submodel that includes $p-1$ parameters, then the LR test statistic on one degree of freedom can be expressed in terms of AIC_0 and AIC_1 as follows (cf. Chapter 6.9.3 in Burnham and Anderson; 2002):

$$T = \text{AIC}_0 - \text{AIC}_1 + 2. \tag{6.3}$$

[1] Similar considerations can be made in the case of nested models differing by the inclusion of more than one parameter.

From Eq. (6.3) one can see that if both models fit the data equally well according to the AIC (i.e., $AIC_0 = AIC_1$), one has $T = 2$. Further, the unrestricted model is chosen over the submodel if its AIC is smaller, corresponding to $AIC_0 - AIC_1 > 0$ and, according to Eq. (6.3), $T > 2$. In contrast, the submodel is chosen if $AIC_0 - AIC_1 < 0$, corresponding to $T < 2$. These considerations show that in the case of two nested models one can also use the value of the LR test statistic to decide which of the models is better in terms of the AIC; if the two models differ only in the inclusion of one parameter, values for the LR test statistic below 2 indicate the superiority of the submodel, values above 2 indicate that the unrestricted model is better, and both models are considered equally good if the LR test statistic takes the value 2. As shown in Section 5.2, bootstrapped LR test statistics systematically deviate from LR test statistics derived from the original data. Due to the correspondence between the LR test statistic and the AIC in the specific setting of nested models it follows that bootstrapped information criteria like the AIC are thus not valid either. These considerations were also made by Wagenmakers et al. (2004). In the context of graphical models, Steck and Jaakkola (2003) proved that bootstrapped information criteria systematically deviate from information criteria derived from original samples.

Although the bootstrapped AIC (6.2) deviates from the AIC (6.1), it is unknown if this discrepancy impacts the decision for a model. One might argue that the discrepancy is not of any practical interest if it does not affect the model choice. To investigate if model choice is affected some experiments using the NHANES data were performed (see Appendix B for information on the NHANES data). With 28 covariates in the NHANES data there are $2^{28} = 268,435,456$ candidate models and, due to computational effort it is not practicable to consider all. One usually considers models that include more than one covariate. But for ease of illustration only 28 models, each arising from the inclusion of exactly one of the variables, are considered. It is investigated which of the models provides the best fit according to the AIC and bootstrapped AIC. Bootstrapped AIC values were computed for 10000 bootstrap samples and an average AIC value was computed.

Figure 6.1 shows the difference between the AIC values computed on the original NHANES sample and the average bootstrapped AIC value. The difference seems to be bigger for models that include more parameters. Though all models have a systematically smaller bootstrapped AIC value, those models incorporating larger numbers of parameters have an exceedingly small AIC value. There are three exceptions: the model featuring *WBCcount*, that for *BMI* and that for *waistcircum*. Note that these are the models with the best model fit according to the AIC. The phenomenon that models incorporating larger numbers of parameters have an exceedingly small AIC value leads to a preferential selection of more complex models. This can be seen when ranking the models according to their average bootstrapped AIC and the AIC obtained for the original sample.

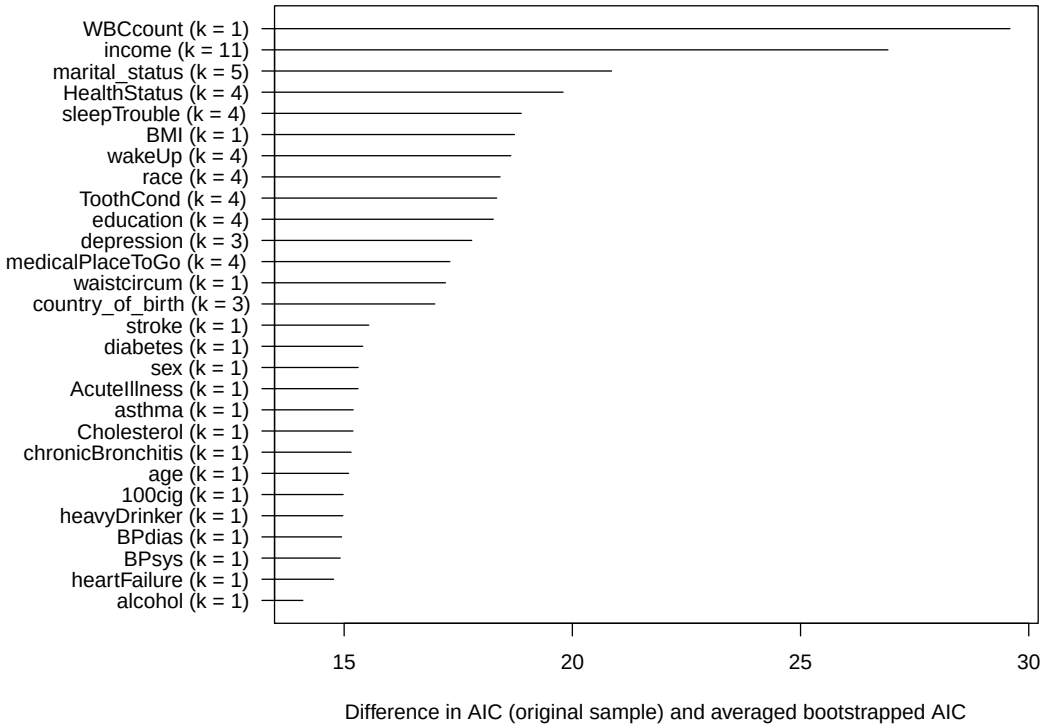

Figure 6.1.: Difference between the AIC value computed on the original sample and the AIC value obtained from averaging over 10000 bootstrapped AIC values for 28 univariate models. The parameter k denotes the number of parameters included in the model for the respective variable.

Figure 6.2 shows the ranking of models by AIC value obtained for the original sample (upper left) and the ranking by the average bootstrapped AIC (upper right). While the top and the bottom of the ranking lists are nearly identical, a number of differences can be observed in the middle: The model which includes $k = 5$ parameters coding marital status is ranked at the 12th position based on the original NHANES sample, while based on bootstrap samples it is ranked 9th (see also Table 6.1). Conversely, the model which includes the variable *sex* ($k = 1$) is ranked 9th based on the original sample but only 12th when AICs were derived from bootstrap samples. Considerable differences in the ranking position can also be observed for the model which includes educational background ($k = 4$). For the original sample this model is ranked only 22nd, while for bootstrap samples it is ranked 17th. Overall, when looking at both rankings, one can see that models which include more parameters seem to obtain higher rankings when ranked by bootstrapped AICs. This applies for the models based on the covariates *wakeUp*, *sleepTrouble*, *marital_status*, *depression*, *education*, or *country_of_birth*. Models which include only one parameter (in addition to the intercept) have lower rankings for bootstrapped AICs (for covariates: *sex*, *AcuteIllness*, *100cig*, *chronicBronchitis*, *age*, *alcohol*, *heartFailure*, *BPsys*, *heavyDrinker*). There are only two exceptions where it is reverse (*Cholesterol* and *race*). These

results strongly suggest that there is a preferential selection of more complex models – i.e., those that include more parameters – when using bootstrapped AICs.

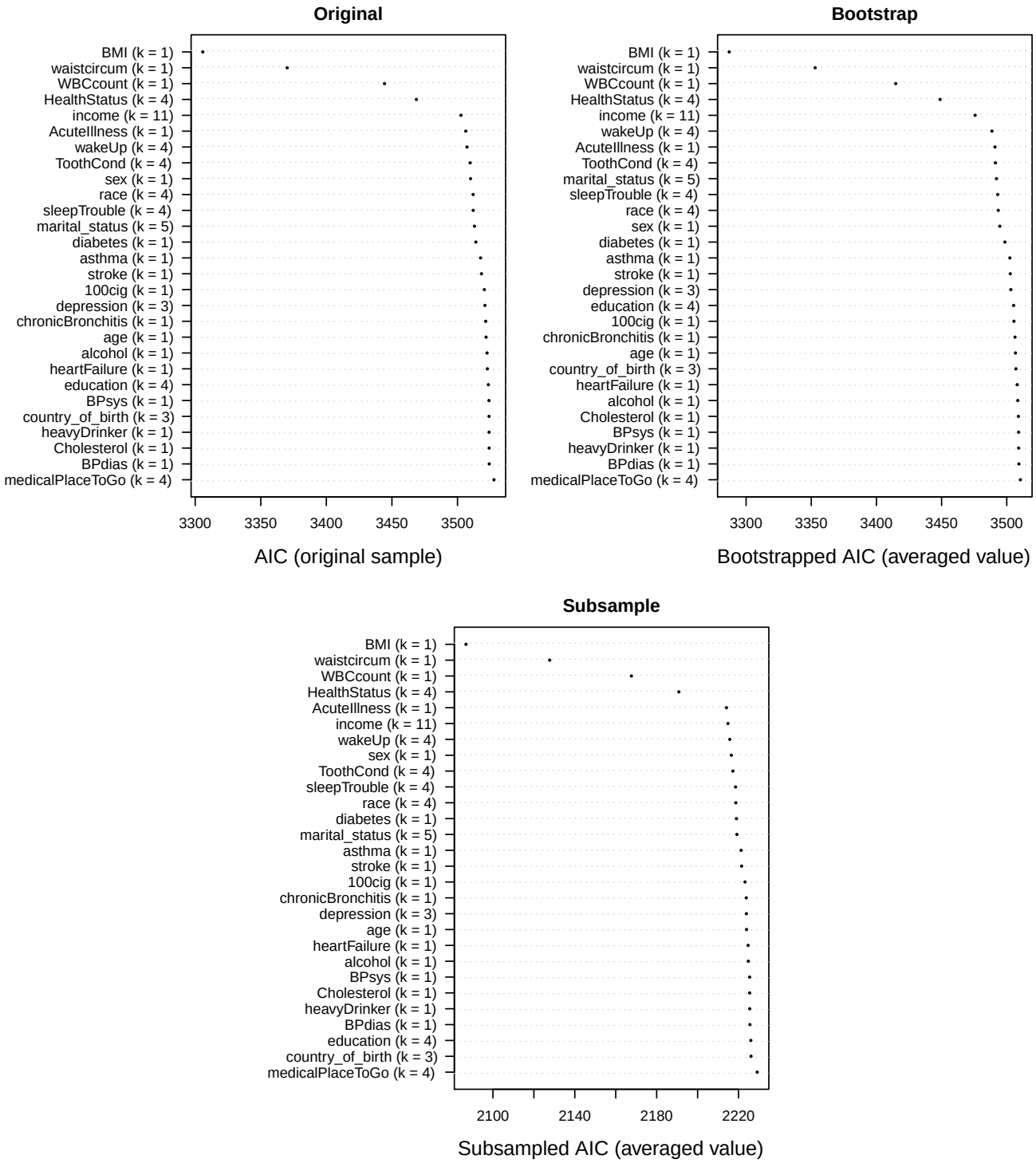

Figure 6.2.: AIC values (in ascending order from top to bottom) obtained for the 28 models (each including exactly one covariate). The parameter k denotes the number of parameters included in the model for the respective variable. Upper left: AIC values derived on the original NHANES sample. Upper right: AIC values obtained from averaging over 10000 bootstrapped AIC values. Lower: AIC values obtained from averaging over 10000 AIC values computed based on subsamples.

Results were also obtained when using subsamples instead of bootstrap samples. Since subsamples contain fewer observations, AIC values obtained for models on subsamples are not comparable to those obtained for the original sample. However, it is interesting to explore if the decision for or against a model is different when the AIC is computed on subsamples instead of the original sample. This can again be seen when sorting the models according to their AIC values (Figure 6.2, lower panel). Indeed there are some characteristic changes in the ordering of the models according to the average AIC obtained from subsamples. But in contrast to the bootstrap, it seems that more complex models (in terms of included parameters) are rather disfavored (see also Table 6.1 and Figure 6.3). This can be explained as follows: From the definition of the AIC in Eq. (6.1) one can see that the AIC is dominated by the penalty term $2p$ (which penalizes the complexity of the model) if the first term $-2\log(L)$ is small, or equivalently, if the likelihood is large. Conversely, the AIC is dominated by the first term, $-2\log(L)$ (which is a measure of the model fit to the data), if the likelihood is small. The likelihood, as a product of n probabilities, becomes automatically smaller with increasing n. As a consequence the likelihood derived from a subsample is larger than the likelihood of the original sample. From these considerations it is clear that for subsamples the AIC is more driven by the penalty term than for the original sample, which leads to the observed phenomenon that more complex models are more disfavored in subsamples than in the original sample. To conclude, AICs obtained from subsamples and original samples do not lead to the same conclusion regarding the choice of optimal models as well.

Application: Gradient boosting

In this section, it is investigated whether there is a preference for more complex models (in terms of included parameters) when constructing models based on bootstrap samples in the special context of gradient boosting (Friedman; 2001; Hothorn et al.; 2010). Gradient boosting has become a popular method in biometrical applications to find sparse models by only making use of relevant predictor variables, which greatly facilitates model interpretation. Briefly, the idea of gradient boosting algorithms is to combine weak learners in an iterative fashion to obtain a strong learner with high prediction accuracy. The prediction accuracy depends highly on the number of iterations, also called the number of boosting steps. With too many boosting steps, many weak learners are constructed and the resulting strong learner might be overfit to the data and thus have poor prediction accuracy on new data. If the number of boosting steps is too small, the number of weak learners might be too small to appropriately model the relationship between the covariates and the response. Thus the number of boosting steps has to be carefully chosen, for example through application of information criteria or internal cross-validation. For

Model complexity	Included variable	Rank for the original sample	Bootstrap rank (diff.)		Subsample rank (diff.)	
$k = 1$	BMI	1	1	(0)	1	(0)
	waistcircum	2	2	(0)	2	(0)
	WBCcount	3	3	(0)	3	(0)
	AcuteIllness	6	7	(-1)	5	$(+1)$
	sex	9	12	(-3)	8	$(+1)$
	diabetes	13	13	(0)	12	$(+1)$
	asthma	14	14	(0)	14	(0)
	stroke	15	15	(0)	15	(0)
	100cig	16	18	(-2)	16	(0)
	chronicBronchitis	18	19	(-1)	17	$(+1)$
	age	19	20	(-1)	19	(0)
	alcohol	20	23	(-3)	21	(-1)
	heartFailure	21	22	(-1)	20	$(+1)$
	BPsys	23	25	(-2)	22	$(+1)$
	heavyDrinker	25	26	(-1)	24	$(+1)$
	Cholesterol	26	24	$(+2)$	23	$(+3)$
	BPdias	27	27	(0)	25	$(+2)$
$k = 3$	depression	17	16	$(+1)$	18	(-1)
	country_of_birth	24	21	$(+3)$	27	(-3)
$k = 4$	HealthStatus	4	4	(0)	4	(0)
	wakeUp	7	6	$(+1)$	7	(0)
	ToothCond	8	8	(0)	9	(-1)
	race	10	11	(-1)	11	(-1)
	sleepTrouble	11	10	$(+1)$	10	$(+1)$
	education	22	17	$(+5)$	26	(-4)
	medicalPlaceToGo	28	28	(0)	28	(0)
$k = 5$	marital_status	12	9	$(+3)$	13	(-1)
$k = 11$	income	5	5	(0)	6	(-1)

Table 6.1.: Model rank by AIC computed for the original sample, by the average bootstrapped AIC, and by the average subsampled AIC. The difference to the rank for original sample is given in brackets for each model. The parameter k denotes the number of parameters included in the model for the respective variable.

more information on gradient boosting algorithms see, for example, Friedman (2001) and Hothorn et al. (2010). The R package mboost was used in all studies presented in this chapter (Hothorn et al.; 2013; Hofner et al.; 2014).

Real data study

The CRP level was the response variable and the 28 variables presented in Table B.1 were considered as candidate predictors in gradient boosting models. Note that, in contrast to the earlier analysis, the association between CRP level and the covariates is now modeled in a multivariate fashion. The AIC was used for choosing the number of boosting steps. Model selection was performed on

- the original NHANES data with $n = 1914$ observations,
- 1000 bootstrap samples, and
- 1000 subsamples.

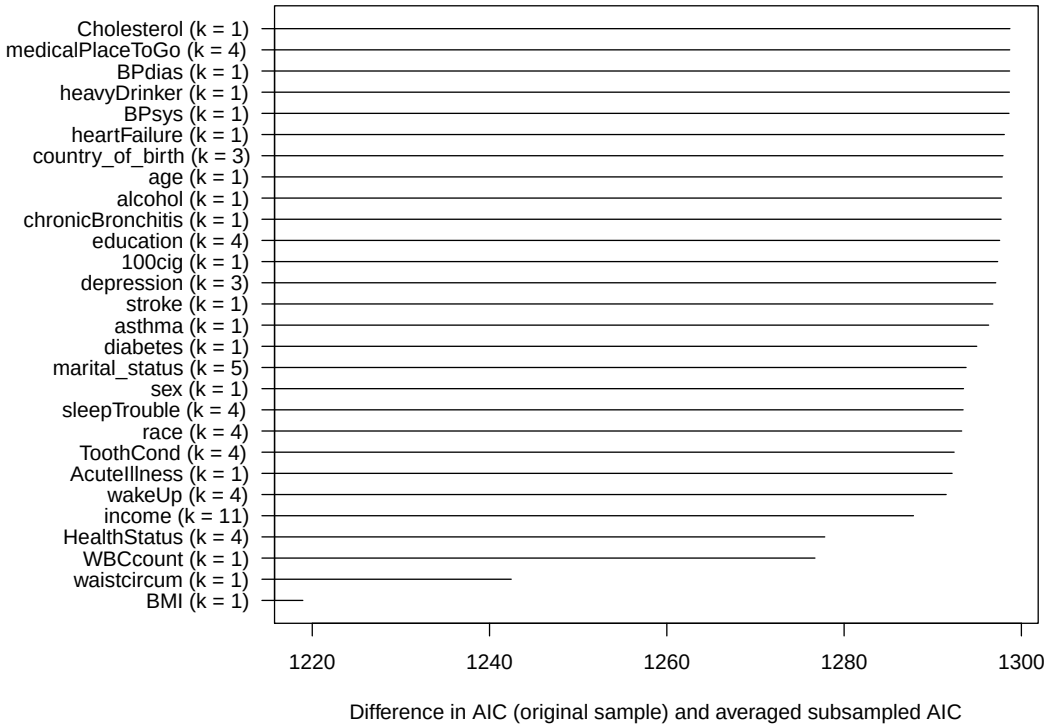

Figure 6.3.: Difference between the AIC value computed on the original sample and the AIC value obtained from averaging over 10000 subsampled AIC values for 28 univariate models. The parameter k denotes the number of parameters included in the model for the respective variable.

For the original NHANES sample, the number of boosting steps for the model with the smallest AIC was 309, the result being a model of 42 parameters (not including the intercept term). When performing tuning parameter selection on bootstrap samples systematically larger values for the number of boosting steps were obtained: in almost all (978 of 1000) bootstrap samples the chosen number of boosting steps was greater than 309 (see left boxplot in Figure 6.4). The mean number of boosting steps in bootstrap samples was 468. The resulting models included a larger number of parameters on average: the average number was 44.3, two parameters more than the model which was obtained for the original NHANES sample. The left panel of Figure 6.5 shows the relative frequency of models with a specific number of parameters. In 68.3% of the bootstrap samples the model included more than 42 parameters, in 24.7% the number of parameters was lower and in 7% the models included exactly 42 parameters.

The same calculations were performed using subsamples instead of bootstrap samples. As expected, sparser models were selected (on average 34.7 parameters) than for the original sample or bootstrap samples (right panel of Figure 6.5). The number of boosting steps (254 on average) was smaller for subsamples, seen in Figure 6.4 (right boxplot).

The models were also evaluated with respect to their predictive accuracy, using the observations that were not drawn into the bootstrap sample and subsample, respectively.

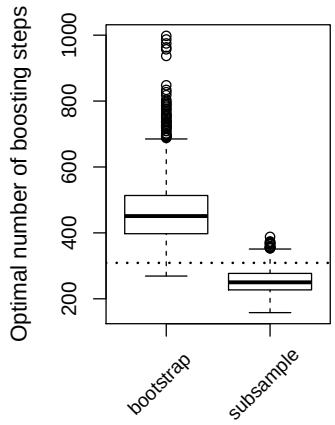

Figure 6.4.: Optimal number of boosting steps selected via AIC in 1000 bootstrap samples and 1000 subsamples of the NHANES data. The dotted horizontal line indicates the chosen number of boosting steps in the original NHANES sample.

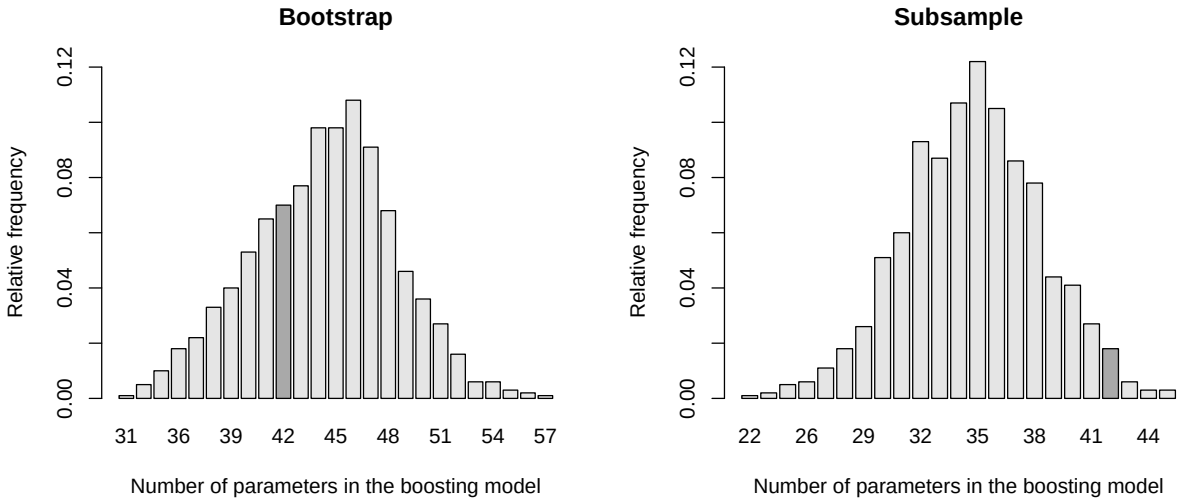

Figure 6.5.: Relative frequency of boosting models (out of 1000) fit on bootstrap samples (left) and on subsamples (right) with specified number of parameters (not including the intercept term). The dark gray bars indicate the number of parameters in the model that was fit on the original NHANES sample.

Although models constructed on subsamples included fewer parameters, their predictive accuracy was comparable to the accuracy of models constructed on bootstrap samples. On average, even a marginally smaller mean squared error was obtained for models fit on subsamples (0.32 compared to 0.33 when using bootstrap samples). This suggests that the additional parameters in the models derived from bootstrap samples do not have any additional predictive value.

Simulation study

Simulation studies with binary response and large numbers of candidate predictors were performed to see if the results differ to those of the NHANES data. The data generating process is the same as that described by Binder and Schumacher (2008) for their simulation study on binary response gradient boosting. Data was simulated for the uncorrelated setting, where $p \in \{200, 1000, 5000\}$ predictors were independently drawn from a standard normal distribution for $n = 100$ observations. The covariate effects are defined as follows:

$$\beta_j = \begin{cases} c_e, & \text{if } j \cdot 200/p \in \{1,3,5,7,9\} \\ -c_e, & \text{if } j \cdot 200/p \in \{2,4,5,6,10\} \\ 0, & \text{otherwise} \end{cases}$$

where $c_e = 1$ (setting with weak effects) and $c_e = 2$ (setting with moderate effects), as per Binder and Schumacher (2008). The binary response, Y_i, for an observation i with covariates $\boldsymbol{x}_i$ was simulated from a Bernoulli distribution with success probability $\pi_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. The AIC was again used to determine the optimal number of boosting steps. The optimal number of boosting steps, the number of included parameters and the accuracy of models were computed on the original data, a bootstrap sample and a subsample. This was repeated 1000 times. Note that, in contrast to the studies with the NHANES data, each bootstrap sample or subsample was drawn from a different original sample.

The accuracy of models was measured on an independent test set of size $n = 10000$ in terms of the Brier score which is defined as

$$BS = \frac{1}{n} \sum_{i=1}^{n} (\hat{\pi}_i - Y_i)^2,$$

with $\hat{\pi}_i$ denoting the predicted probability that Y_i equals 1 for an observation i with covariates $\boldsymbol{x}_i$.

Figure 6.6 shows the results on the optimal number of boosting steps (upper row), on the number of parameters included in the respective model (middle row), and on prediction accuracy (lower row). The results are shown for the setting with weak effects; those for the setting with moderate effects are comparable and thus not shown. Overall, the prediction accuracies of models derived from subsamples and bootstrap samples are very similar. As with the real data study, the prediction accuracy of models derived from subsamples was even marginally better, although models fit on bootstrap samples included slightly more parameters.

As far as the selection of optimal numbers of boosting steps is concerned, a marginally higher number of boosting steps was chosen in bootstrap samples than in original sam-

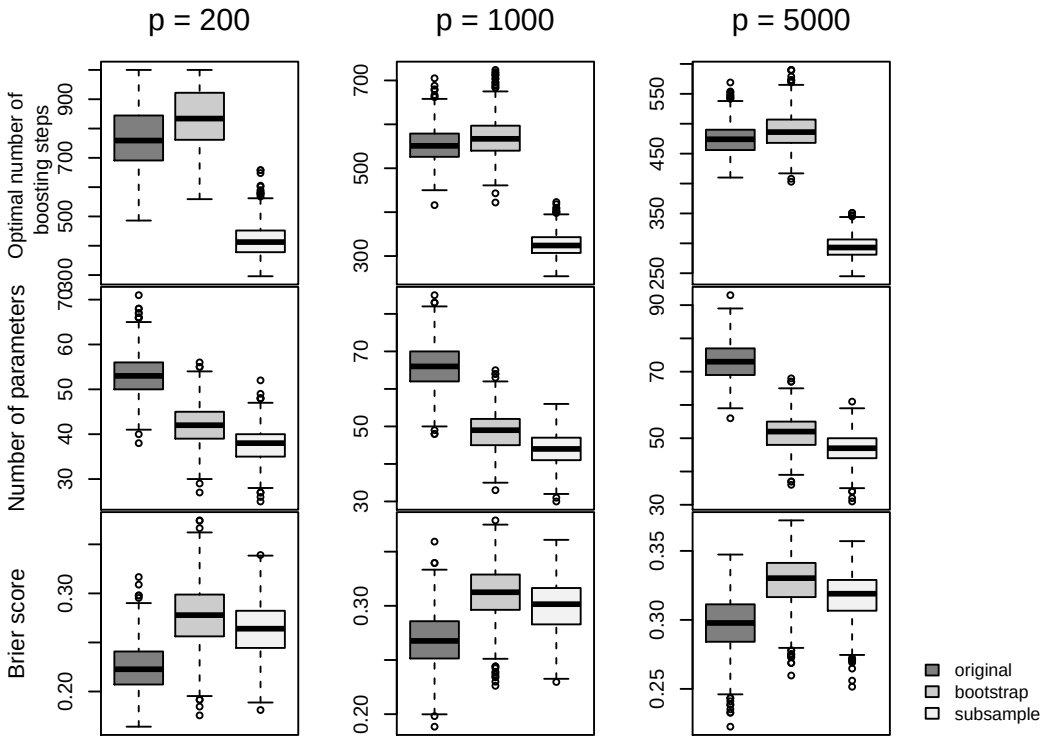

Figure 6.6.: Optimal number of boosting steps (upper row), the number of parameters in the corresponding model (middle row), and the prediction accuracy measured by the Brier score (lower row). Results are shown for models selected via AIC for binary response gradient boosting in 1000 original samples, bootstrap samples and subsamples.

ples. In subsamples in contrast, a substantially smaller number of boosting steps was chosen than for original samples. The difference in the optimal number of boosting steps was much larger between original samples and subsamples than between original samples and bootstrap samples. However, the marginally larger number of boosting steps chosen in bootstrap samples did not lead to the inclusion of more parameters in the resulting models. In fact, the models fit on original samples included more parameters. Further inspection revealed that the relationship between the number of boosting steps and the number of parameters included in a model is a different one in original samples and in bootstrap samples or subsamples. Figure 6.7 shows the average number of parameters included in a model (over 1000 original samples, 1000 bootstrap samples and 1000 subsamples, resp.) for a specified number of boosting steps. For bootstrap samples and subsamples the relationship between the number of boosting steps and the number of parameters seems to be very similar. For numbers of boosting steps above 100 the models fit on bootstrap samples or subsamples include fewer parameters compared to models fit on original samples. The largest differences are seen in the setting with the largest number of candidate predictors ($p = 5000$; right column). For numbers of boosting steps below 100 all methods include an approximately equal number of parameters. Note that

in each boosting step, either a new variable enters the boosting model, or a parameter of a variable which is already in the model, is updated. In the present studies it was seen that in original samples a new parameter entered the boosting models more frequently than in bootstrap samples or subsamples. In bootstrap samples and subsamples, in contrast, the parameters of variables that were already included in a model, were updated more frequently than in original samples. Further research is needed to understand the reasons. However, these findings explain why the larger number of boosting steps chosen based on the AIC on bootstrap samples does not necessarily lead to a larger number of parameters contained in the respective model.

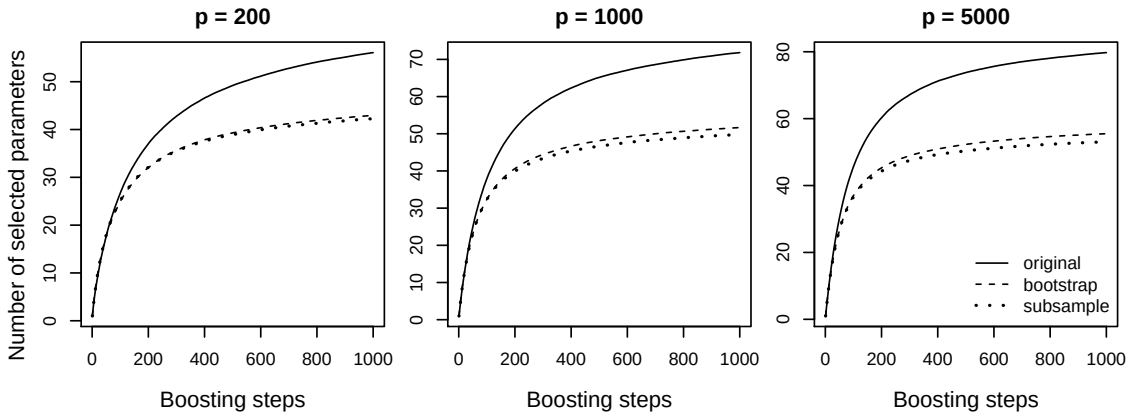

Figure 6.7.: Relationship between the number of boosting steps and the number of parameters contained in a model for settings with $n = 100$ observations and $p = 200$ (left), $p = 1000$ (middle) and $p = 5000$ (right) candidate predictors. The average number of parameters for a specific number of boosting steps was derived from 1000 original samples, bootstrap samples and subsamples, respectively.

The simulation results seem to contradict the results obtained for the real data set, where a larger number of boosting steps resulted in overcomplex models on bootstrap samples. Note that the real data set included 1914 observations and 67 parameters (from 28 candidate predictors). The real data study is thus very different from the considered simulated scenarios, in which the number of candidate predictors is much larger than the number of observations. When performing additional simulation studies, in which the number of observations exceeds the number of candidate predictors, the results are in line with those for the real data set. Figure 6.8 shows the average number of parameters included in a model for a specific number of boosting steps in settings with $p = 50, n = 100$ (left panel), $p = 100, n = 500$ (middle panel) and $p = 50, n = 1000$ (right panel). The larger the number of observations and the smaller the number of candidate predictors, the more similar the relationship between the number of boosting steps and the number of selected parameters is for the different sampling schemes. In these settings, again,

larger numbers of boosting steps were chosen based on bootstrap samples. In contrast to the earlier simulation studies, the resulting models included more parameters than the models which were fit on original samples (results not shown).

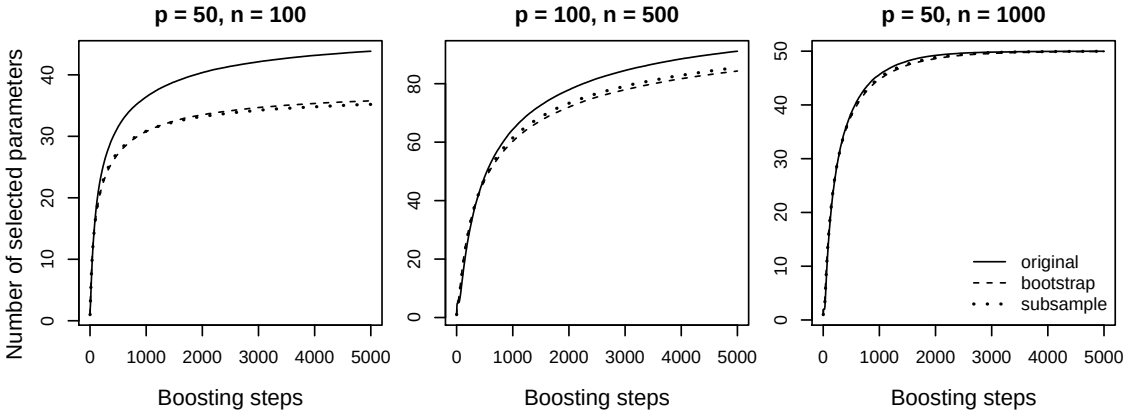

Figure 6.8.: Relationship between the number of boosting steps and the number of parameters contained in a model for settings with $p = 50, n = 100$ (left), $p = 100, n = 500$ (middle) and $p = 50, n = 1000$ (right). The average number of parameters for a specific number of boosting steps was derived from 1000 original samples, bootstrap samples and subsamples, respectively.

To conclude, neither models derived from bootstrap samples nor models derived from subsamples are of similar complexity as models derived from original samples. Whether models derived from bootstrap samples are either more complex than models derived from original samples or less complex depends on the considered data setting. For gradient boosting methods the theory that models fit on bootstrap models are always more complex does thus not apply. Subsampling often resulted in models that included far fewer parameters than the models derived from original samples. Therefore they do not reflect the complexity of models fit on the original data. However, if one aims at finding models that have a good predictive ability, then subsampling might be preferable over bootstrapping as models had marginally better predictive ability if they were fit on subsamples.

6.2. Tuning parameter selection through data splitting approaches

An alternative strategy to the use of the AIC is choosing a value for the tuning parameter that yields models which have high predictive ability. The predictive ability of models has to be assessed on an independent test data set for a grid of candidate values. To obtain unbiased estimates of the models' predictive accuracies, data splitting approaches

in which the data is split into a training set and a test set have to be applied. Note that the data splitting approaches are now applied *within the model building step* and not within the model evaluation step. There are different data splitting procedures such as the bootstrap which is handled in this thesis. Note that, after having found an optimal value for the tuning parameter, the bootstrap is used for the model evaluation step. But for tuning parameter selection within the model building step, cross-validation is considered in this section.

The steps for selecting an optimal value for a tuning parameter through the use of k-fold cross-validation are outlined in the following:

1. *Initialization*: Set l to 1.
2. *Data splitting*: Randomly split the observations of a sample into k sets of equal size.
3. *Tuning model building*: Let S_l contain the indices of observations from the l-th set. Use all observations except for those in S_l (*training set*) to fit models for each considered value of a tuning parameter.
4. *Tuning model evaluation*: Use the observations in S_l (*test set*) to measure the models' accuracies using a pre-defined evaluation criterion (e.g., the mean squared error in the case of metric responses).
5. *Iteration*: If $l < k$ increment l and repeat steps 3 to 5.
6. *Tuning parameter selection*: Average the accuracies of all models which are based on the same value for the tuning parameter, and select the value for the tuning parameter which yields the largest mean prediction accuracy.

It is recommended to repeat the steps 1 to 5 several times to obtain more stable results (see, e.g., Braga-Neto and Dougherty; 2004).

Note that in this section the sample on which cross-validation is performed is either an original sample (i.e., a sample drawn from the true distribution), a bootstrap sample or a subsample. The same observation can be contained several times in a bootstrap sample. Thus when performing cross-validation on a bootstrap sample, the same observation may be present in the training set and in the test set. If a model is evaluated on observations that were already used for fitting the model, more complex models show a better prediction accuracy on these data. However, these have poor predictive accuracy on new data. The use of subsampling might solve this problem since each observation of the original data is contained in a subsample no more than once. A different solution might be to prevent an overlap of training and test sets. Hothorn et al. (2005), for example, propose deleting the observations from the test set that are also present in the training set. In this chapter a different approach for preventing an overlap is considered, in which all duplications of the same observation in a bootstrap sample are regarded as one unit, and the

units – instead of the observations – are randomly split into k sets, in which each set contains an equal number of units. Note that the k sets are then not necessarily of the same size. Gradient boosting methods are an example where data splitting approaches, such as cross-validation, are performed on bootstrap samples for selecting tuning parameters.

Application: Gradient boosting

Binder and Schumacher (2008) investigated cross-validation on bootstrap samples to select the optimal number of boosting steps. Their simulation results consistently show that the number of boosting steps is considerably larger when performing tuning parameter selection on bootstrap samples compared to original samples. The consequence of the considerably high number of boosting steps was overcomplex models with decreased accuracy. However, it was not clear what exactly has led to this overcomplexity in their studies. The real data based study and the simulation study presented in this section (i) investigate if the preference for overcomplex models is induced by the overlap between training and test set or has a different cause, (ii) quantify the extent of overcomplexity, and (iii) give hints which method(s) most closely resemble the models obtained from original data and should thus be used.

Real data study

The studies based on the NHANES data in Section 6.1, where the effect of various factors on the CRP level was modeled, were replicated, but this time using cross-validation to select an optimal number of boosting steps. The number of boosting steps which minimizes the cross-validated mean squared error was considered optimal. Model selection was performed on

- the original NHANES data with $n = 1914$ observations,
- 1000 bootstrap samples allowing training and test sets to overlap,
- 1000 bootstrap samples not allowing training and test sets to overlap, and
- 1000 subsamples.

In the studies 10 runs of 5-fold cross-validation were performed.

For the original NHANES sample the chosen number of boosting steps was 104. The corresponding model (fit on the whole NHANES data) included 19 parameters plus the intercept term.

The numbers of boosting steps chosen in bootstrap samples (with and without allowing training and test sets from 5-fold cross-validation to overlap) and in subsamples are shown in Figure 6.9. In 89.5% of the bootstrap samples where training and test sets may

overlap, the chosen number of boosting steps was greater than 104; the average number of boosting steps was 227.2, and thus is much larger than the number of boosting steps selected in the original NHANES sample. The numbers of parameters included in the corresponding bootstrap models are shown in Figure 6.10 (upper left panel). Models derived from bootstrap samples included systematically more parameters than the original model, and the average number of included parameters was 30.3 (compared to only 19 parameters for the NHANES sample). As already seen in the studies of Binder and Schumacher (2008), a considerably higher model complexity – in terms of parameters included in a model – is obtained when performing cross-validation on bootstrap samples (where training and test sets may overlap) compared to original samples.

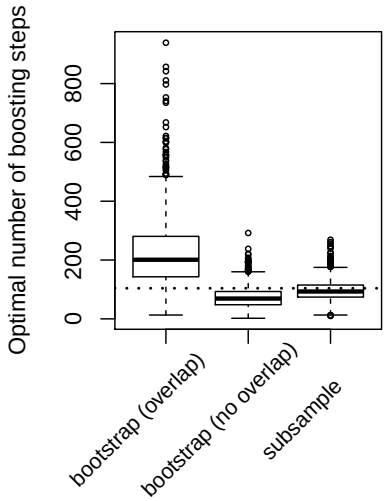

Figure 6.9.: Optimal number of boosting steps selected via 5-fold cross-validation in 1000 bootstrap samples (with and without allowing training and test sets from repeated 5-fold cross-validation to overlap) and 1000 subsamples of the NHANES data. The dotted horizontal line indicates the chosen number of boosting steps in the original NHANES data.

In contrast to that, there is no tendency toward more complex models when performing cross-validation on bootstrap samples with the restriction that training and test sets do not overlap. The contrary is the case: both the number of boosting steps and the number of parameters included in the models are systematically smaller than for the original sample (see Figures 6.9 and 6.10). This is also seen from Table 6.2 which shows the average and median numbers of boosting steps and included parameters. The average number of boosting steps was 72.1 (compared to 104 in the original sample) and the average number of included parameters was 13.8 (compared to 19) for the bootstrap approach in which training and test sets do not overlap.

The results obtained from bootstrap samples in which training and test sets do not overlap, come close to the results obtained for subsamples. Since subsamples are drawn such that the number of unique observations is approximately the same as for bootstrap

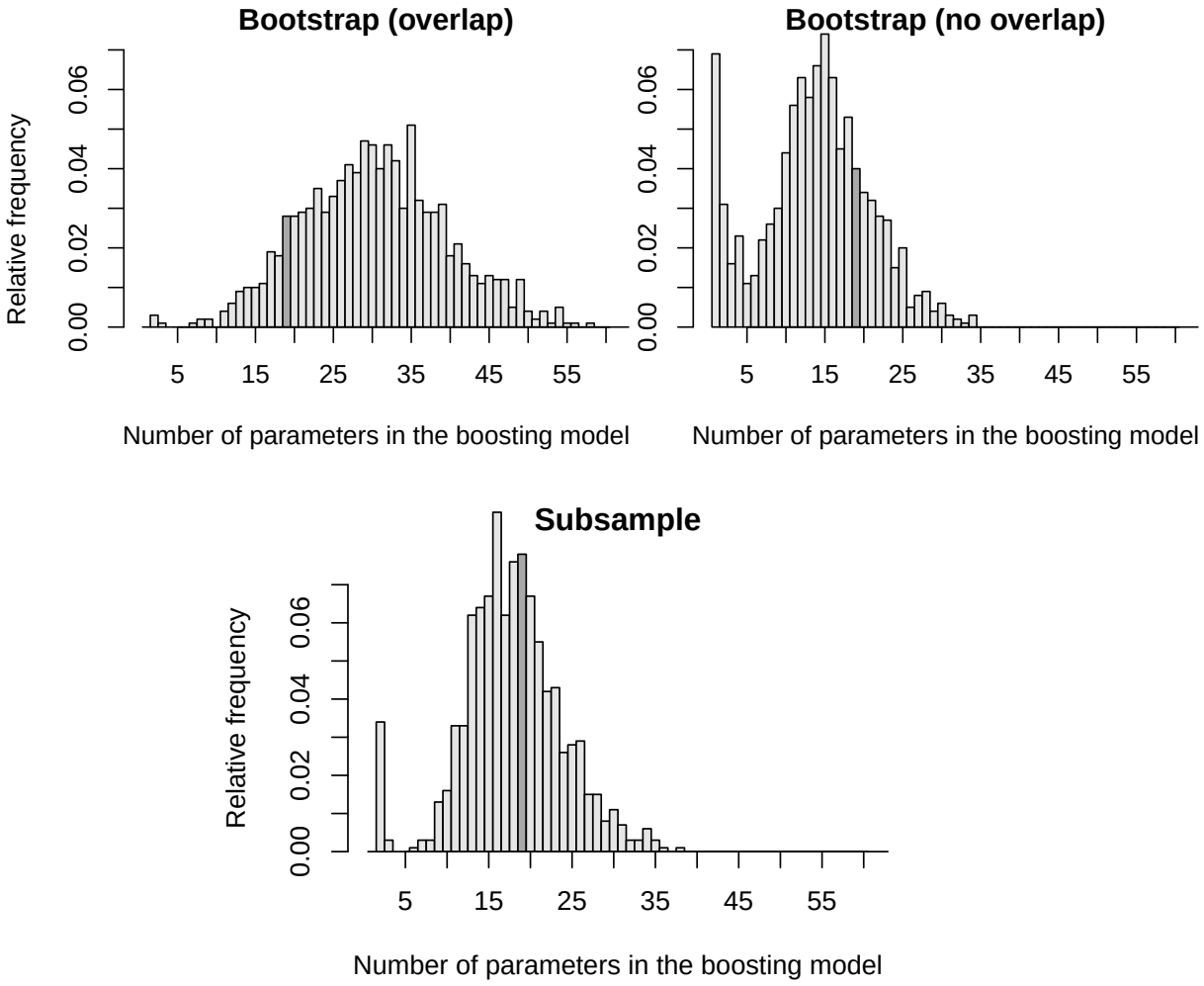

Figure 6.10.: Relative frequency of boosting models (out of 1000) fitted on bootstrap samples (with and without allowing training and test sets from 5-fold cross-validation to overlap) and on subsamples with specified number of parameters (not including the intercept term). The dark gray bars indicate the number of parameters in the model that was fit on the original NHANES sample.

samples (i.e., 63.2% of the original sample), one might hypothesize that the information content in a bootstrap sample and in a subsample is approximately the same, thus leading to similar models. There are, though, some differences between the two approaches. While for subsamples the model complexity is very close to that of the original NHANES sample, the models for bootstrap samples (with restriction that training and test sets do not overlap) are less complex. The duplication of single observations in a sample obviously affects boosting algorithms to some extent. Alternatively, the observed differences are not due to the duplicated observations, but to the ratio of training and test set sizes, which is not constant over all k training and test set splits. Further studies are needed to investigate the reasons which lead to the differences. However, the studies make it clear that the higher complexity of gradient boosting models results from the overlap of

	Optimal no. of boosting steps		No. of included parameters	
Original	104		19	
	Mean	Median	Mean	Median
Bootstrap (overlap)	227.185	201	30.253	30
Bootstrap (no overlap)	72.075	69	13.758	14
Subsample	96.071	93	17.867	18

Table 6.2.: Mean and median numbers of boosting steps and parameters included in the resulting boosting models fit on 1000 bootstrap samples (with and without allowing training and test sets from 5-fold cross-validation to overlap) and on 1000 subsamples. The numbers for the original sample are also shown.

training and tests sets when performing cross-validation on bootstrap samples.

Although models derived from bootstrap samples where training and test sets may overlap, included far more parameters than models derived from the other two sampling approaches, there was hardly any difference regarding the models' predictive accuracy. The average mean squared error was 0.32 for subsamples and 0.33 for both bootstrap approaches.

To conclude, the subsampling procedure yielded the best results in the considered real data application. The models derived from subsamples included almost as many parameters as the model obtained from the original sample, thus reflecting the complexity of the original model, and, in addition, they had a similar, and even marginally smaller mean squared error than models fit on bootstrap samples.

Simulation study

The simulation design is adopted from Binder and Schumacher (2008) who first encountered the overcomplexity of models built from bootstrap samples when parameter tuning is performed using cross-validation. Their simulation design with binary response has been described in Section 6.1. Model selection was performed on

- 1000 original samples, each including $n = 100$ observations,
- 1000 bootstrap samples allowing training and test sets to overlap,
- 1000 bootstrap samples not allowing training and test sets to overlap, and
- 1000 subsamples.

Note that each bootstrap sample or subsample was drawn from a different original sample. Ten runs of 5-fold cross-validation were performed.

The results for the setting with weak effects ($c_e = 1$) are shown in Figure 6.11. Those for the setting with moderate effects ($c_e = 2$) were comparable and are thus not shown. The studies show that with the classical bootstrap approach (where training and test sets may

overlap) too complex models are promoted. As seen in the upper row of Figure 6.11, the number of selected boosting steps is much larger than that obtained for original samples. The number of parameters in a model is substantially larger, too, and the difference between the numbers for original and bootstrap samples increases with increasing numbers of candidate predictors (middle row). Performing cross-validation on bootstrap samples for choosing the optimal number of boosting steps is thus not recommended because the resulting models are much more complex than models fit on original samples, especially in settings with large numbers of candidate predictors.

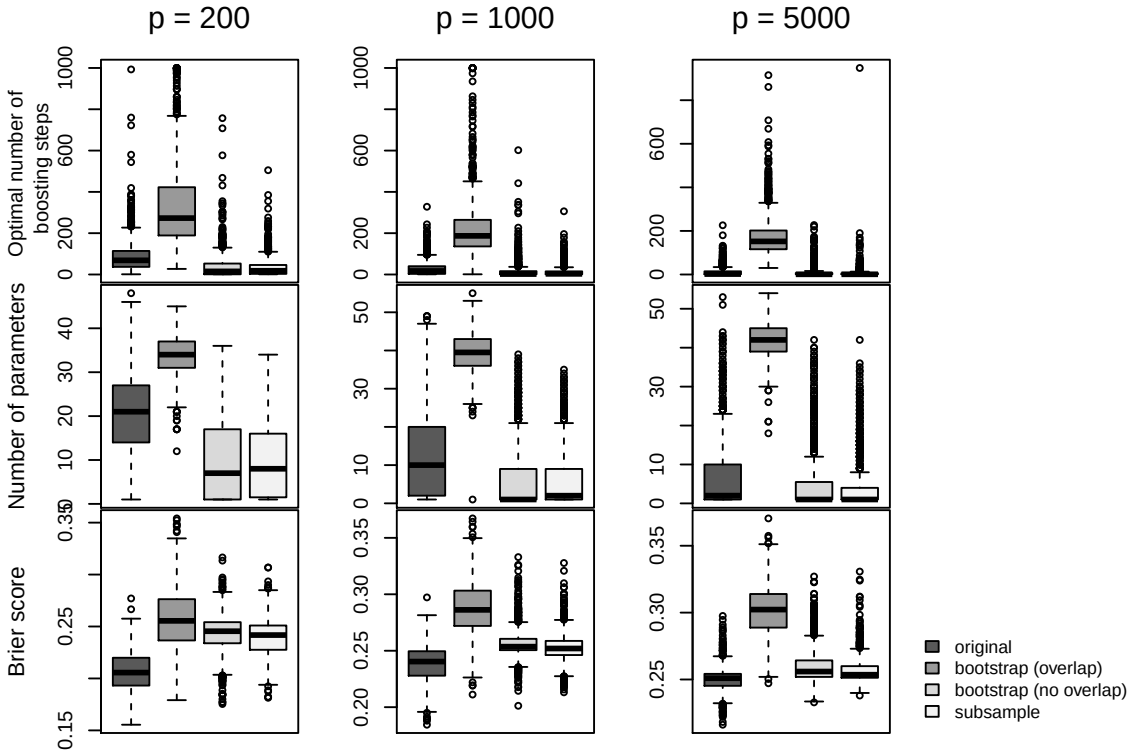

Figure 6.11.: Optimal number of boosting steps (upper row), the number of parameters in the corresponding model (middle row), and the prediction accuracy which was measured through the Brier score (lower row). Results are shown for models selected via 5-fold cross-validation for binary response gradient boosting in 1000 original samples, bootstrap samples (with and without allowing training and test sets from repeated 5-fold cross-validation to overlap), and subsamples.

The results obtained for subsampling and those obtained for the bootstrap approach, in which training and test sets do not overlap, are similar. This is in line with the findings of the real data study. With subsampling and the bootstrap approach that prevents an overlap, sparser models are chosen compared to models derived on original samples. Both the number of boosting steps and the number of parameters are smaller for these two sampling approaches than for original samples. The difference is smaller for the settings with larger numbers of candidate predictors. In the settings with $p = 5000$, the model complexities for the modified bootstrap approach and for subsampling approximate the

complexity of models fit on original samples well. However, in the settings with $p = 200$, models fit on original samples are much more complex.

The prediction accuracy is worst for models fit on bootstrap samples in which training and test sets overlap (cf. Figure 6.11, lower row). The accuracy of models fit on bootstrap samples without overlap and subsamples is better. In the setting with $p = 200$ the prediction accuracies of models obtained through subsampling and both bootstrap approaches are comparable. In the other two settings, subsampling and the bootstrap approach that prevents an overlap of training and test sets distinctly outperform the classical bootstrap approach, especially in the setting with $p = 5000$. This suggests that the overcomplexity induced by the bootstrap might result in models that are less accurate. Further, the studies suggest that this problem becomes more relevant in settings with large numbers of predictor variables.

6.3. Discussion

Based on the results presented in Chapter 5 it was shown that bootstrapped information criteria do not represent what would be obtained on the original data. Models of different complexity are the consequence, as seen in studies with gradient boosting models. In settings where the number of predictor variables exceeded the number of observations, boosting models were less complex when fit on bootstrap samples, while in low-dimensional settings boosting models were more complex. Future theoretical investigations on the bootstrapped AIC in gradient boosting models are needed to study the reasons for this.

The studies in this chapter suggest that using cross-validation for tuning parameter selection in bootstrap samples is much more problematic than using information criteria since the respective boosting models included a substantially larger number of parameters than boosting models derived for original samples. The overcomplexity was shown to be likely attributable to the overlap between training and test sets when cross-validation was performed on bootstrap samples. Subsampling might be used to avoid this problem. However, the models derived from subsamples included fewer parameters than models derived from original samples and thus do not reflect the complexity of models obtained for the original samples, either. A modified bootstrap strategy in which training and test sets do not overlap, was investigated. The results of this modified bootstrap strategy were similar to those obtained for subsampling, and the modified bootstrap strategy also yielded models which are systematically less complex than models derived from original samples. Future research might aim at developing alternative strategies that provide models which are similar to models obtained for original sam-

ples. If the aim is to obtain accurate prediction models, however, the studies suggest that subsampling should be preferred over the bootstrap.

Another interesting finding was that the numbers of boosting steps chosen in original samples and in bootstrap samples cannot be compared. In some studies presented in this chapter, the number of boosting steps was larger for bootstrap samples than for original samples but the respective bootstrap models included less parameters than the models derived from the original samples. Thus one cannot use the selected number of boosting steps as a substitute for the number of parameters in a model, for comparing models on original samples with models on bootstrap samples (or subsamples, resp.). If it is used wrong conclusions regarding the models' complexity might be obtained.

Finally, it should be noted that the chosen boosting model depends highly on the method for tuning parameter selection (cross-validation or AIC). The studies show that boosting models for original samples are much more complex if the AIC is used. The AIC is computed from the degrees of freedom which are unknown for gradient boosting models and have to be estimated. The available AIC estimates, however, underestimate the degrees of freedom (Hastie; 2007). The consequence is that systematically too many boosting steps are chosen when using the AIC as a criterion. In the considered studies, the numbers of parameters included in models for original samples were always substantially different for the two tuning parameter selection strategies. The difference increased with increasing numbers of candidate predictors. While in the (low-dimensional) real data set there was no difference in the models' prediction accuracies, in the (high-dimensional) simulation study the models selected based on AIC had worse performance than the models selected based on cross-validation, which were much less complex. In light of the results from the simulation studies, the general recommendation to prefer cross-validation over the AIC for tuning parameter selection in gradient boosting models, can be approved (see Mayr et al.; 2012, for recommendations on the selection of the optimal number of boosting steps).

7. Conclusion and outlook

This thesis focused on resampling approaches with special emphasis on bootstrap-based procedures. The first and second parts of this thesis dealt with one specific bootstrap-based procedure, namely random forests (RF).

RF are well investigated for classification and regression tasks. However, there is a lack of studies dealing with the application of RF for ordinal regression. Chapter 3 addressed this issue. Existing and new approaches were investigated for both prediction and variable selection. The studies were based on the RF methodology of Hothorn, Hornik and Zeileis (2006). Besides classification and regression trees, this RF version implements so-called ordinal regression trees. When constructing ordinal regression trees, the ordinal response is treated like a metric response, but contrary to regression trees, ordinal regression trees provide class predictions instead of real-valued predictions. Ordinal regression trees have not been tested so far. Extensive studies shown in Chapter 3 indicate that prediction performance of RF consisting of ordinal regression trees and RF consisting of classification trees is very similar, and that there is hardly any improvement in prediction performance when making use of the ordering of the response levels. In contrast to that, variable ranking can be improved when using novel variable importance measures (VIMs) which make use of the ordering. These new VIMs are special types of permutation VIMs. They are computed from the out-of-bag ranked probability score or from the out-of-bag mean absolute error, respectively. Studies show that variable rankings are most reliable when using the novel VIMs in combination with ordinal regression trees.

A drawback of VIMs is that there is no natural cutoff for importance scores that can be used to differentiate between important and non-important variables. When using RF's VIMs every researcher thus faces the problem which variables to select from the ranking list. Permutation-testing approaches have been developed for addressing this problem. While for low-dimensional settings approaches based on permutation tests can be applied, for high-dimensional settings these approaches are computationally very demanding – if not infeasible. In Chapter 4 a computationally fast heuristic testing procedure for high-dimensional data was proposed. This testing procedure makes use of the observed non-positive importance scores to approximate the distribution of importance scores for non-relevant variables (null distribution). For each variable a p-value can then easily be derived based on this approximation of the null distribution. The new test-

ing approach is based on the assumption of a null distribution symmetric around zero. This assumption is, however, not met when deriving the importance scores based on the classical permutation VIM. Therefore, a modified version of the permutation VIM was developed. This modified version is not computed from the out-of-bag observations but is based on cross-validation procedures. The null distribution based on this modified VIM was empirically shown to be symmetric around zero and thus fulfills the requirements of the novel testing procedure. In the studies with high-dimensional data the new testing approach controlled the type I error and had at least the same statistical power as the permutation-test-based approach of Altmann et al. (2010).

In some studies the modified VIM had a slightly better discriminative ability than the classical VIM. The reasons for this are unknown and future studies are needed to investigate if at all – or in which settings – the modified VIM is superior to the classical VIM. Such studies are valuable in that they detect deficiencies of the classical VIM and aid the development of improved VIMs.

In the studies presented in Chapter 4 only settings with categorical response were considered. Further studies are needed to examine the performance of the novel approach in settings with metric response or in settings with ordinal response considered in Chapter 3. The performance of the novel testing approach for metric response was studied by Celik (2015) in his master thesis which was initiated and supervised by myself. Overall, the novel approach showed good performance in the studies of Celik (2015). In some studies the type I error was marginally increased, but the increase in type I error observed in the studies may possibly not be regarded as relevant in practice. However, from theoretical considerations it is expected that the testing approach might fail in specific settings. In the studies of Nicodemus et al. (2010), it was shown that, even if none of the correlated variables is associated with the response, the importance score distribution of highly correlated predictor variables lies in the positive range. This cannot be repaired through the use of the novel permutation VIM: As indicated in the studies of Celik (2015), the null distribution for the novel VIM is in the positive range, too. This could result in increased type I error of the novel testing approach. However, for high-dimensional settings the distribution was far less shifted on the x-axis in positive direction than for low-dimensional data settings (Celik; 2015). This might explain that the testing approach (almost) preserved the type I error in the studies of Celik (2015). But one cannot exclude the possibility of increased type I error in future settings including highly correlated predictor variables. Nicodemus et al. (2010) showed that the null distribution is not shifted in the positive range when using the conditional VIM of Strobl et al. (2008) in combination with conditional inference trees. Therefore, a promising solution to the problem is to derive the novel cross-validation based VIM from the conditional VIM and not from the

classical VIM. This derivation is straightforward as the cross-validation based VIM can be derived from any arbitrary VIM that makes use of out-of-bag observations for evaluating the trees' performance. It is expected that there is no increased type I error when using a cross-validation based VIM which is derived from the conditional VIM.

However, an important problem persists when using RF's VIMs for variable selection. This problem concerns the interpretation of the selected variables. Due to the non-parametric nature of RF there is no easy interpretation of the variables in RF. In comparison with model-based approaches, which allow for an easy – but simplistic – interpretation of the effect of variables through the use of regression coefficients, the interpretation of the effect of variables in RF is awkward. Approaches, such as partial dependence plots, have been developed and might be a good starting point. Future work is needed to address this challenging problem.

Besides its use in ensemble methods, such as RF, the bootstrap is widely used in biometry to solve problems that are difficult to address based on asymptotic theory. With the introduction of the bootstrap in 1979 more and more approaches which are based on the bootstrap have been developed. Recently some approaches have been suggested where hypothesis tests or model selection through information criteria or data splitting approaches are performed on bootstrap samples as if they were original samples. The third and forth part of this thesis dealt with problems related to the use of hypothesis testing or model selection strategies on bootstrap samples.

In Chapter 5 it was shown that the p-values of tests performed on bootstrap samples as if they were the original samples do not represent what would be obtained on the original data, and that there is increased type I error for tests performed on bootstrap samples. These findings were based on theoretical and empirical investigations using the Z-test and the likelihood ratio test as examples. The practical impact on three procedures was assessed and promising alternative strategies, such as subsampling, were considered. The first application makes use of the bootstrap for investigating the stability of the results obtained from stepwise model selection procedures that implement hypothesis testing for deciding which variables to include in a model or eliminate from a model (Chen and George; 1985; Altman and Andersen; 1989; Sauerbrei and Schumacher; 1992). It was shown that performing stepwise procedures on bootstrap samples results in models that include more parameters compared to models obtained from original samples. Moreover, categorical variables with many categories are preferentially selected in a model. The second considered procedure aims to generate stable variable ranking lists based on bootstrapped p-values (Mukherjee et al.; 2003). It was shown that these procedures fail if there are variables of different scales for the same reason. In the third application bootstrapped p-values are used for assessing the variability of p-values (Bollen and

Stine; 1992). In most of the considered simulation settings the variability estimated from bootstrapped p-values was systematically too large or too small and thus did not reflect the p-value variability of original samples.

Bootstrapping is commonly used for the estimation of the error of a prediction modeling strategy. Many statistical methods involve tuning parameters. Optimal values for these parameters are usually found through the application of information criteria or cross-validation procedures. Chapter 6 investigated the use of bootstrapped information criteria and cross-validation for tuning parameter selection in bootstrap samples with a focus on gradient boosting models. When using the AIC for tuning parameter selection in gradient boosting models, studies showed that it depends on the specific data setting, whether boosting models fit on bootstrap samples are either more complex or less complex than boosting models fit on original samples. In low-dimensional data settings, boosting models fit on bootstrap samples were rather more complex, while in high-dimensional settings they were less complex. In contrast, boosting models fit on bootstrap samples were always more complex when cross-validation was used for tuning parameter selection. It was shown that the higher complexity of boosting models results from the overlap of training and tests sets when performing cross-validation on bootstrap samples.

Subsampling resulted in boosting models which were less complex than models fit on original samples. This was the case for both tuning parameter selection through AIC and cross-validation. The complexity of boosting models fit on subsamples thus does not represent the complexity of boosting models fit on original samples, either. However, if the aim is to find sparse boosting models that have high predictive accuracy, subsampling should be preferred over the bootstrap. In all studies, models fit on subsamples had slightly better prediction accuracies than models fit on bootstrap samples, although the differences were marginal. This suggests that the additional parameters included in models constructed based on bootstrap samples do not have any additional predictive value.

The studies presented in this thesis showed that one cannot directly apply any procedure to bootstrap samples as if they were the original sample. The list of bootstrap approaches mentioned in this thesis is with certainty not complete. Applied researchers should be careful when using approaches to problems in which hypothesis tests or model selection through information criteria or data splitting approaches are performed based on a bootstrap sample. Depending on the considered method and on the aim of a study, the bootstrap might be very useful or it might lead to wrong conclusions. If no investigations exist that indicate the reliability of a bootstrap approach, simulation studies are a helpful investigative tool.

A. Steps for deriving the random forests prediction rule

Step 1: Deriving prediction rules on the training data

This section gives details on classifiers that were fit on the training data in the anaerobic blood culture samples. The analyses were performed using Bioconductor package CMA (Slawski et al.; 2008).

Parameter tuning was done using internal 3-fold cross-validation. The grids of candidate parameter values were chosen as the default grids in CMA version 1.10.0. Parameters that were not tuned were chosen as the default values in CMA version 1.10.0 if not explicitly specified here.

Variable selection was performed before fitting classifiers using the following methods:

- ranking of variables by their p-values using a t-test,
- ranking of variables by their p-values using Limma (Smyth; 2005), or
- ranking of variables by their variable importance score via random forest's Gini importance measure (RF VIM).

Dimension reduction was performed before fitting classifiers using partial least squares (PLS).

Table A.1 gives an overview of the fitted classifiers.

Step 2: Selecting the best prediction rule

The accuracies of the prediction rules were computed in an unbiased way using 5-fold cross-validation (repeated 100 times), while preserving response class distributions within all folds (stratified sampling). The prediction rules were compared with respect to their cross-validated prediction errors. Figures A.1 and A.2 show the distribution of error rates over the 100×5 repetitions (prediction rules are in the same order as in Table A.1). The

minimal error rate is obtained for the random forests prediction rule when using only the 10 highest ranked variables according to the RF VIM. This prediction rule was considered the best.

Step 3: Fitting the random forests prediction rule on the whole training data

The next step was to build the random forests prediction rule with the 10 top ranked variables (by RF VIM) using the whole training data. This prediction rule might be used for predicting future independent data. First the 10 top ranked variables by the Gini VIM were identified from the whole training data. These variables are H_2, 34, 35, 35*, 36, 64, 64*, 66, 76*, 80*, whereas the asterisk indicates that compounds were measured by chemical ionization using xenon instead of mercury as primary ion. The tuning parameters (i.e., the number of randomly selected variables at each split and the minimal number of observations in a node) were selected through 3-fold cross-validation performed on the training data. The optimal value for the number of randomly selected variables at each split was 3 and the optimal number for the minimum size of nodes was 7. These values were used to build the final random forests prediction rule consisting of 1000 trees based on the whole training data.

Step 4: Validating the random forests prediction rule

The random forests prediction rule was validated using an independent validation data that arose from randomly splitting the original data into a training and a validation set (ratio 2:1). The error rate, sensitivity and specificity, the ROC and the area under the curve are shown in Section 2.2.2.

Method	CMA option	Tuned parameters	Variable selection and/or dimension reduction	Parameters differing from default
Componentwise boosting with binomial loss function	`compBoostCMA`	no. of boosting steps	-	-
Componentwise boosting with exponential loss function	`compBoostCMA`	no. of boosting steps	-	-
Componentwise boosting with quadratic loss function	`compBoostCMA`	no. of boosting steps	-	-
Tree-based gradient boosting with Bernoulli loss function	`gbmCMA`	no. of trees	-	-
Tree-based gradient boosting with exponential loss function	`gbmCMA`	no. of trees	-	-
Random forests	`rfCMA`	no. of randomly selected variables at each split; minimal number of observations in a node	-	-
Random forests	`pls_rfCMA`	no. of components	PLS	-
Random forests with $p \in \{10, 20, \ldots, 100\}$ top variables from variable ranking	`rfCMA`	no. of randomly selected variables at each split; minimal number of observations in a node	RF VIM	-
Support vector machines with linear kernel	`svmCMA`	cost parameter	-	-
Support vector machines with polynomial kernel	`svmCMA`	cost parameter; degree of the polynomial kernel	-	-
Support vector machines with radial kernel	`svmCMA`	cost parameter; width of the radial basis function kernel	-	-
L1 penalized logistic regression	`LassoCMA`	penalization parameter	-	-
L2 penalized logistic regression	`plrCMA`	penalization parameter	-	-
Elastic net penalized logistic regression	`ElasticNetCMA`	penalization parameters for L1 and L2 norm	-	-
k nearest neighbors	`knnCMA`	no. of nearest neighbors	-	-
k nearest neighbors with $p \in \{10, 20, \ldots, 100\}$ top variables from variable ranking	`knnCMA`	no. of nearest neighbors	t-test	-
k nearest neighbors with $p \in \{10, 20, \ldots, 100\}$ top variables from variable ranking	`knnCMA`	no. of nearest neighbors	Limma	-
Probabilistic k nearest neighbors	`pknnCMA`	no. of nearest neighbors	-	beta = 0.00015
Probabilistic k nearest neighbors with $p \in \{10, 20, \ldots, 100\}$ top variables from variable ranking	`pknnCMA`	no. of nearest neighbors	t-test	beta = 0.00015
Probabilistic k nearest neighbors with $p \in \{10, 20, \ldots, 100\}$ top variables from variable ranking	`pknnCMA`	no. of nearest neighbors	Limma	beta = 0.00015

Feed forward neural networks with $p \in \{10, 20, \ldots, 100\}$ top variables from variable ranking	`nnetCMA`	weight decay parameter	t-test	-
Feed forward neural networks with $p \in \{10, 20, \ldots, 100\}$ top variables from variable ranking	`nnetCMA`	weight decay parameter	Limma	-
Linear discriminant analysis with $p \in \{2, 4, 6, \ldots, 40\}$ top variables from variable ranking	`ldaCMA`	-	t-test	-
Linear discriminant analysis with $p \in \{2, 4, 6, \ldots, 40\}$ top variables from variable ranking	`ldaCMA`	-	Limma	-
Linear discriminant analysis	`pls_ldaCMA`	no. of components	PLS	-
Linear discriminant analysis with $p \in \{10, 20, 30, \ldots, 100\}$ top variables from variable ranking	`pls_ldaCMA`	-	t-test, PLS	-
Linear discriminant analysis with $p \in \{10, 20, 30, \ldots, 100\}$ top variables from variable ranking	`pls_ldaCMA`	-	Limma, PLS	-
Fisher's linear discriminant with $p \in \{10, 20, 30, \ldots, 60\}$ top variables from variable ranking	`fdaCMA`	-	t-test	-
Fisher's linear discriminant with $p \in \{10, 20, 30, \ldots, 60\}$ top variables from variable ranking	`fdaCMA`	-	Limma	-
Diagonal linear discriminant analysis	`dldaCMA`	-	-	-
Diagonal linear discriminant analysis with $p \in \{5, 10, 15, \ldots, 100\}$ top variables from variable ranking	`dldaCMA`	-	t-test	-
Diagonal linear discriminant analysis with $p \in \{5, 10, 15, \ldots, 100\}$ top variables from variable ranking	`dldaCMA`	-	Limma	-
Quadratic discriminant analysis with $p \in \{2, 4, 6, \ldots, 18\}$ top variables from variable ranking	`qdaCMA`	-	t-test	-
Quadratic discriminant analysis with $p \in \{2, 4, 6, \ldots, 18\}$ top variables from variable ranking	`qdaCMA`	-	Limma	-

Table A.1.: Prediction rules fit on training data.

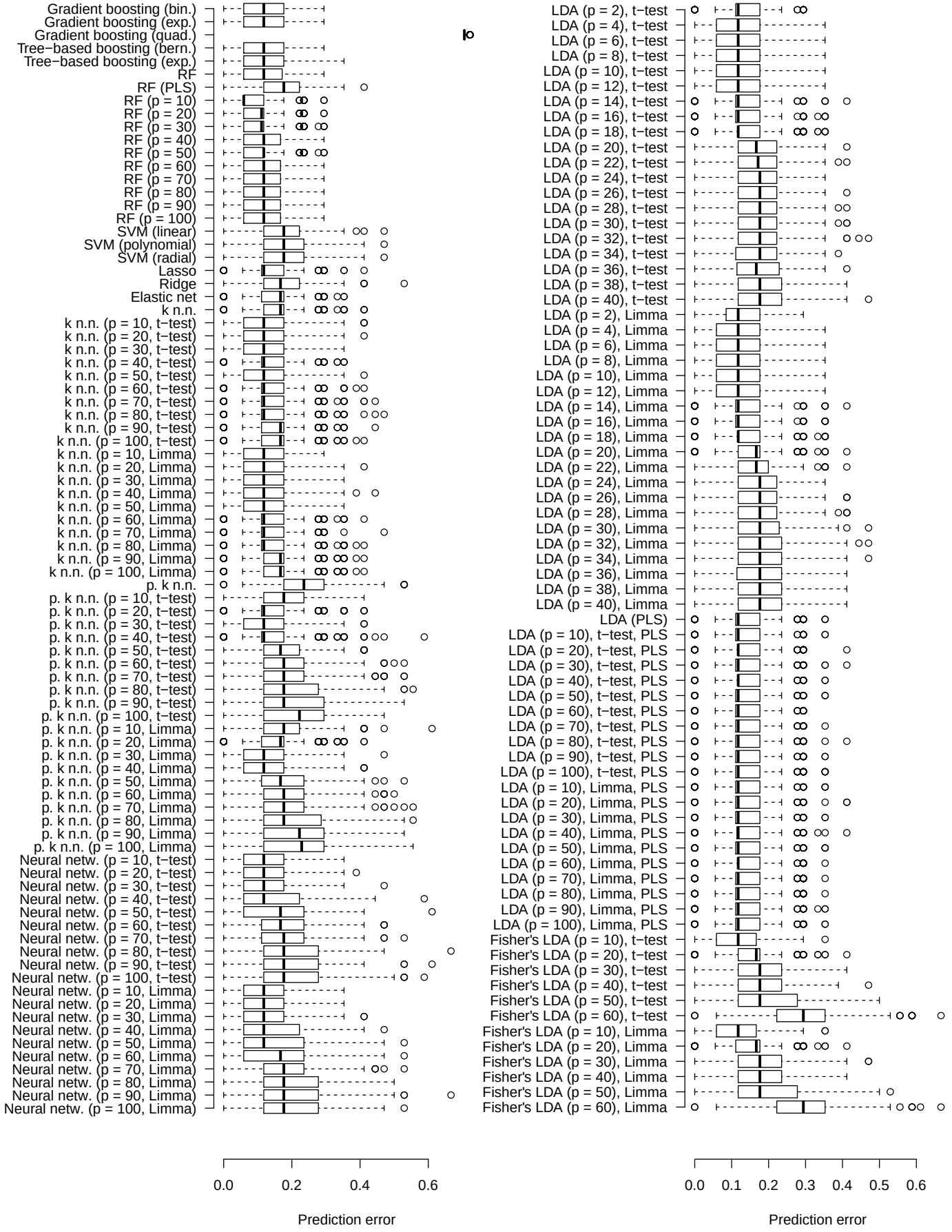

Figure A.1.: Comparison of prediction rules constructed based on the training data. Parameter p denotes the number of preselected variables. RF: random forests; SVM: support vector machines, k n.n.: k nearest neighbors; p. k n.n.: probabilistic k nearest neighbors; LDA: linear discriminant analysis.

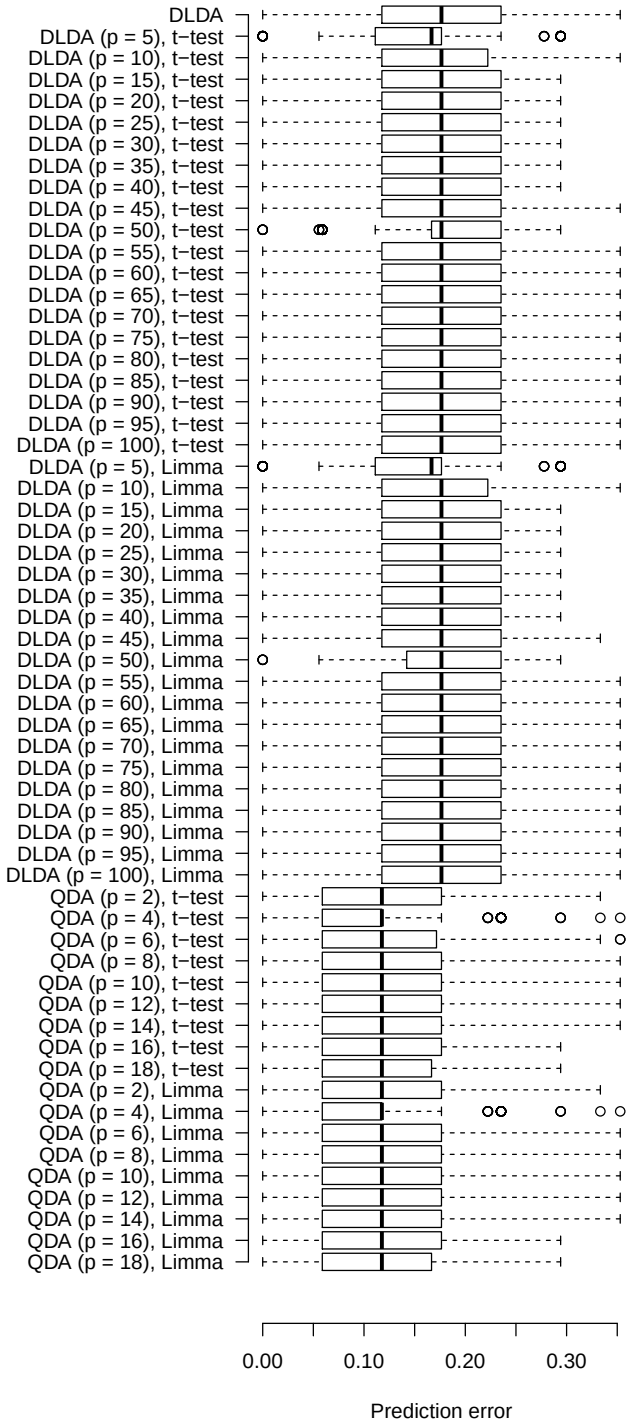

Figure A.2.: Comparison of prediction rules constructed based on the training data (cont.). Parameter p denotes the number of preselected variables. DLDA: diagonal linear discriminant analysis; QDA: quadratic discriminant analysis.

B. NHANES data

The data considered in Chapters 3, 5 and 6 is from the 2007-2008 cycle of the National Health and Nutrition Examination Survey (NHANES) (National Center for Health Statistics; 2012) which is maintained by the Centers for Disease Control and Prevention. NHANES is designed as a series of cross-sectional surveys conducted in the US population. The data are freely available from the institution's homepage or from the Interuniversity Consortium for Political and Social Research. The considered data set comprises a total of $n = 1914$ subjects. For the studies in Chapters 5 and 6, the level of high-sensitive C-reactive protein (CRP) was used as the response. The CRP is a plasma protein involved in the acute phase response during inflammatory states (Black et al.; 2004). Besides CRP there are 28 variables that include information on the medical facility to which the subject most often goes, the subject's sex, age, body mass index, waist circumference, race, country of birth, education, marital status, income, smoking history, information on alcohol consumption, as well as laboratory values (white blood cell counts, systolic and diastolic blood pressure, cholesterol level) and health-related conditions including asthma, diabetes, history of stroke, history of heart failure, history of chronic bronchitis, history of any acute illnesses, history of alcohol abuse, self-rated general health, tooth condition, depressive mood, sleeping abnormalities (consisting of items on falling asleep and waking during the night). Many of the variables were obtained from interviews with the study persons. The corresponding interview questions, the abbreviations for the variables used in Chapters 5 and 6, and the measurement units of the variables are given in Table B.1.

Abbreviation	Interview question / description	Categories / units
race	Recode of reported race and ethnicity information	Mexican American Other Hispanic Non-Hispanic White Non-Hispanic Black Other Race, including Multi-Racial
country_of_birth	In what country (were you/was sample person) born?	50 US States or Washington, DC Mexico Other Spanish Speaking Country Other Non-Spanish Speaking Country
education	What is the highest grade or level of school (you have/sample person has) completed or the highest degree (you have/she/he has) received?	less than 9th up to 11th high school some college graduate
marital_status	Marital status	married widowed divorced separated never married living with partner
HealthStatus	Would you say (your/sample person's) health in general is ...	excellent very good good fair poor
depression	Over the last 2 weeks, how often have you been bothered by the following problems: little interest or pleasure in doing things? Would you say ...	not at all several days over half the days nearly every day
ToothCond	Now I have some questions about the condition of your teeth and gums. How would you describe the condition of (your/sample person's) teeth? Would you say ...	excellent very good good fair poor
sleepTrouble	In the past month, how often did (you/sample person) have trouble falling asleep?	never rarely sometimes often almost always

Continued on next page

Continued from previous page

wakeUp	In the past month, how often did (you/sample person) wake up during the night and had trouble getting back to sleep?	never rarely sometimes often almost always
medicalPlaceToGo	What kind of place (do you/does sample person) go to most often: is it a clinic, doctor's office, emergency room, or some other place?	clinic doctor's office hospital emergency hospital outpatient other
income	Total household income (reported as a range value in dollars)	under $5k $5k - under $10k $10k - under $15k $15k - under $20k $20k - under $25k $25k - under $35k $35k - under $45k $45k - under $55k $55k - under $65k $65k - under $75k $75k - under $100k over $100k
AcuteIllness	Did (you/sample person) have a head cold or chest cold that started during the last 30 days? *or* Did (you/sample person) have flu, pneumonia, or ear infections that started during those 30 days? *or* Did (you/sample person) have a stomach or intestinal illness with vomiting or diarrhea that started during those 30 days?	yes no
100cig	(Have you/Has sample person) smoked at least 100 cigarettes in (your/his/her) entire life?	yes no
diabetes	(Have you/Has sample person) ever been told by a doctor or health professional that (you have/(he/she/sample person) has) diabetes or sugar diabetes?	yes no
asthma	Has a doctor or other health professional ever told (you/sample person) that (you/she/he) have/has asthma?	yes no
heartFailure	Has a doctor or other health professional ever told (you/sample person) that (you/she/he) had congestive heart failure?	yes no
stroke	Has a doctor or other health professional ever told (you/sample person) that (you/she/he) had a stroke?	yes no
chronicBronchitis	Has a doctor or other health professional ever told (you/sample person) that (you/she/he) had chronic bronchitis?	yes no

Continued on next page

Continued from previous page

heavyDrinker	Was there ever a time or times in (your/sample person's) life when (you/he/she) drank 5 or more drinks of any kind of alcoholic beverage almost every day?	yes no
waistcircum	Circumference of waist	cm
Cholesterol	Cholesterol level	mg/dL
WBCcount	White blood cell count	1k cells/μL
BPsys	Systolic blood pressure	mmHg
BPdias	Diastolic blood pressure	mmHg
age	Age	years
BMI	Body mass index	kg/m^2
alcohol	Alcohol consume	units

Table B.1.: Variables and corresponding interview questions or descriptions for the NHANES data.

C. Additional results

C.1. Random forests for ordinal responses

Simulation settings

Additional studies were conducted, in which the number of variables was increased to $p = 1015$. The number of observations was $n = 200$. The response was generated according to a mixture of two proportional odds models as described in Section 3.3.3. Among the 1015 variables $X_1, \ldots, X_{15}$ had an effect on at least one mixture component (see Section 3.3.3 for their coefficient values), while the remaining 1000 variables had no effect and their coefficients for both mixture components were zero. Model parameters and simulation settings were the same as for the simulations described in Chapter 3.

For settings with correlations, $\boldsymbol{x}_i, i = 1, \ldots, 200$, were drawn from $\mathrm{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_p)$ (now $p = 1015$) with block diagonal covariance matrix

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} \boldsymbol{A}_{\text{signal}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{A}_{\text{noise}_{1,1}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{A}_{\text{noise}_{1,2}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_{1,20}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_{2,1}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_{5,20}} \end{bmatrix}.$$

The first block matrix $\boldsymbol{A}_{\text{signal}} \in \mathbb{R}^{(15\times 15)}$ determined the correlations among the signal predictors $X_1, \ldots, X_{15}$. It was defined as $\boldsymbol{A}_{\text{signal}} = (a_{ij})$ with

$$a_{ij} = \begin{cases} 1, & i = j \\ 0.8, & i \neq j; i, j \in \{1, 3, 6, 8, 11, 13\} \\ 0, & \text{otherwise} \end{cases}$$

in this way generating uncorrelated and also strongly correlated signal predictors. The

matrices $\boldsymbol{A}_{\text{noise}_{i,j}} \in \mathbb{R}^{(10\times 10)}$ for $i = 1, \ldots, 5$ and $j = 1, \ldots, 20$ were given by

$$\boldsymbol{A}_{\text{noise}_{i,j}} = \begin{bmatrix} 1 & \rho_i & \ldots & \rho_i \\ \rho_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_i \\ \rho_i & \ldots & \rho_i & 1 \end{bmatrix}$$

and determined correlations among a set of 10 noise predictor variables with $\rho_1 = 0.8$, $\rho_2 = 0.6$, $\rho_3 = 0.4$, $\rho_4 = 0.2$ and $\rho_5 = 0$.

The RF parameter setting described in Section 3.3.3 was chosen. However, since at most 15 of 1015 predictor variables had an effect, a large *mtry* value of 100 was chosen for the present studies.

Results

Figures C.1 – C.4 show the performance ratio for *RF ordinal* versus *RF classification*. Prediction accuracy was evaluated using the ranked probability score (Figure C.1), the error rate (Figure C.2), the class-specific error rates averaged over all classes (Figure C.3) and a generalization of the area under the curve to multiple ordered classes (Figure C.4). The latter is the average over $k - 1$ area under the curves; the l-th area under the curve is given by

$$AUC_l = \frac{1}{\sum_{i=1}^{l} n_i \sum_{j=l+1}^{k} n_j} \sum_{Y_i \leq l} \sum_{Y_j > l} I(\hat{Y}_i < \hat{Y}_j) + 0.5 I(\hat{Y}_i = \hat{Y}_j),$$

with n_r denoting the number of observations in classes $r = 1, \ldots, k$, Y_i and $\hat{Y}_i$ denoting the observed and predicted responses, respectively, for observations $i = 1, \ldots, n$, and $l \in \{1, \ldots, k-1\}$ (see, e.g., Waegeman et al.; 2008).

In contrast to the ranked probability score and the classical error rate, the class-specific error rate and the area under the curve do not give the same weight to each observation, but they account for the fact that response classes may be unbalanced.

The results are similar for all considered prediction accuracy measures. Overall, the prediction performance of *RF ordinal* is slightly better than that of *RF classification*.

The performance of the four considered VIMs is shown in Figures C.5, C.6 and C.7 for the simulation settings with 9, 6 and 3 ordered response classes. The VIMs based on the ranked probability score, mean absolute error and mean squared error are clearly better than the VIM based on the error rate, especially in settings with larger numbers of response classes. The best predictor rankings were obtained when using these VIMs in combination with ordinal regression trees, although in most settings the differences to rankings obtained from classification trees were marginal.

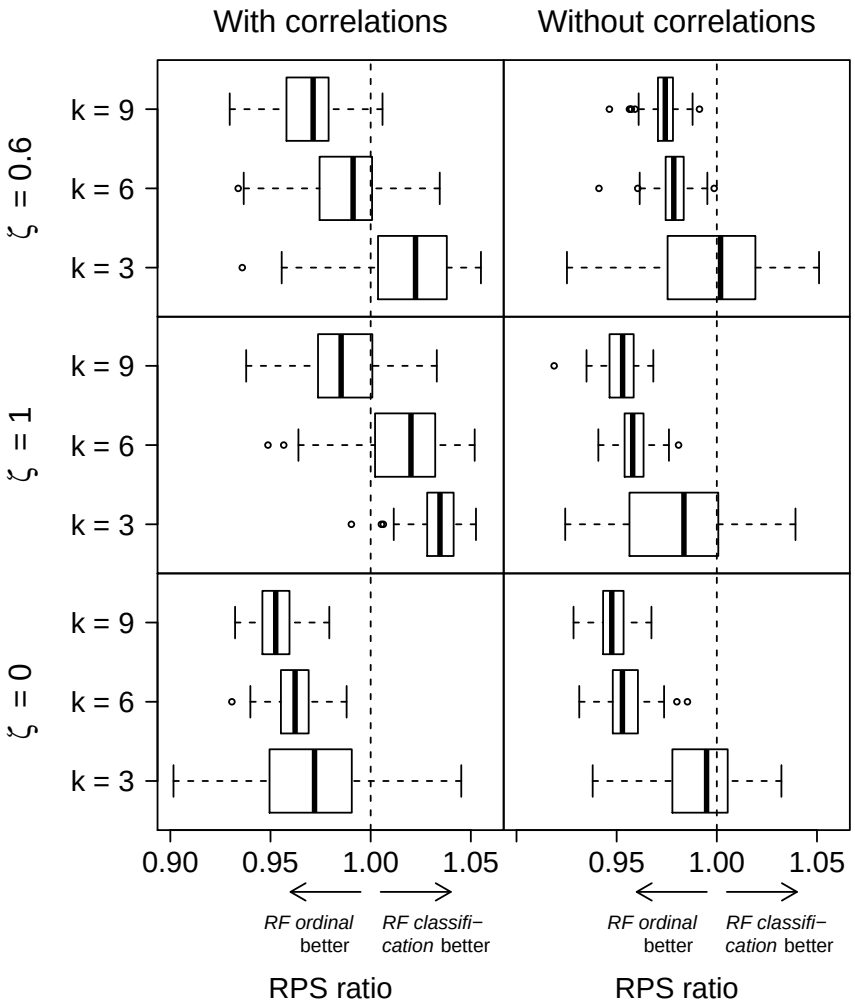

Figure C.1.: Performance ratio for *RF ordinal* versus *RF classification* for simulated data with $p = 1015$ predictor variables for $n = 200$ observations. A ratio of the ranked probability scores (RPS) below 1 indicates a better prediction accuracy of *RF ordinal* and a ratio above 1 indicates a better prediction accuracy of *RF classification*. Data was generated from a mixture of proportional odds models with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row). Data was generated for $k \in \{3, 6, 9\}$ ordered response levels and for settings in which predictors correlate (left column) and in which all predictors are uncorrelated (right column).

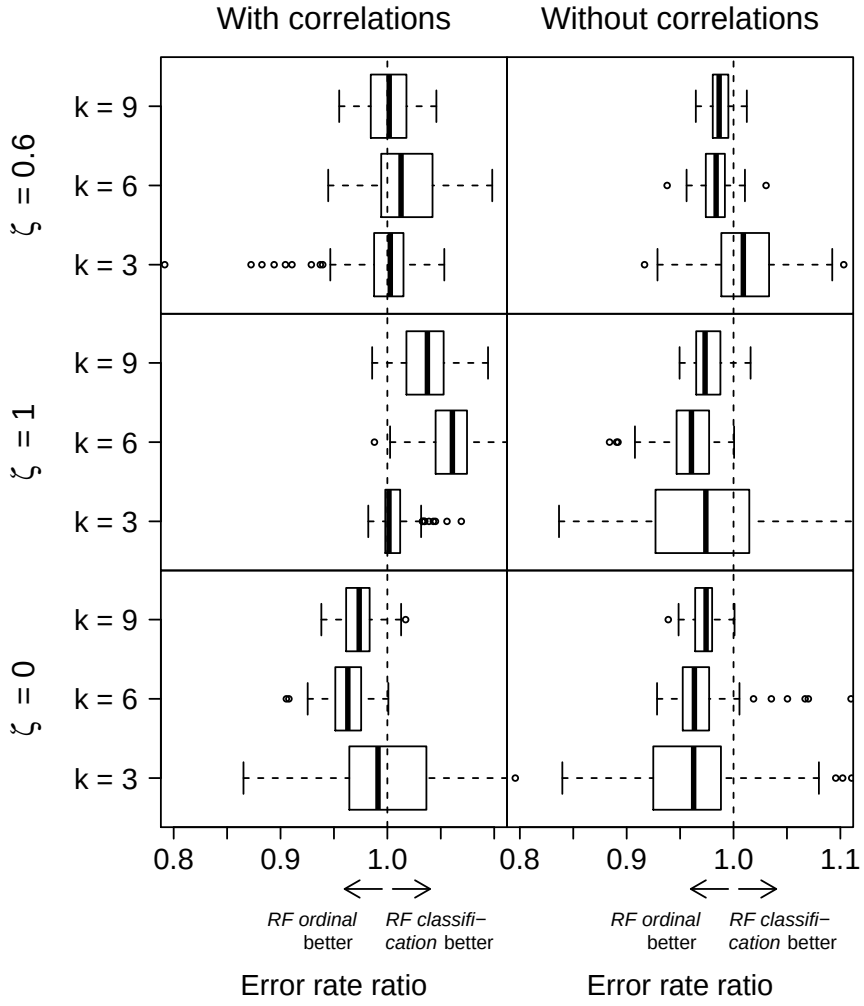

Figure C.2.: Performance ratio for *RF ordinal* versus *RF classification* for simulated data with $p = 1015$ predictor variables for $n = 200$ observations. A ratio of the error rate below 1 indicates a better prediction accuracy of *RF ordinal* and a ratio above 1 indicates a better prediction accuracy of *RF classification*. Data was generated from a mixture of proportional odds models with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row). Data was generated for $k \in \{3, 6, 9\}$ ordered response levels and for settings in which predictors correlate (left column) and in which all predictors are uncorrelated (right column).

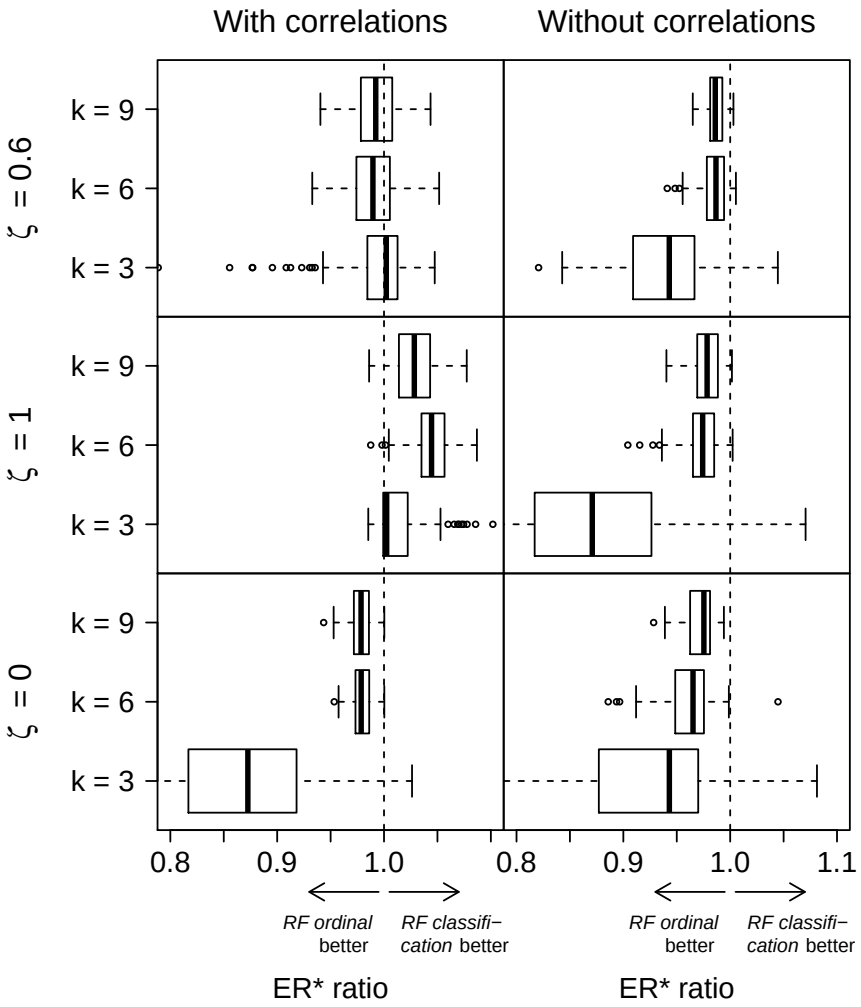

Figure C.3.: Performance ratio for *RF ordinal* versus *RF classification* for simulated data with $p = 1015$ predictor variables for $n = 200$ observations. A ratio of the average over class-specific error rates (ER*) below 1 indicates a better prediction accuracy of *RF ordinal* and a ratio above 1 indicates a better prediction accuracy of *RF classification*. Data was generated from a mixture of proportional odds models with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row). Data was generated for $k \in \{3, 6, 9\}$ ordered response levels and for settings in which predictors correlate (left column) and in which all predictors are uncorrelated (right column).

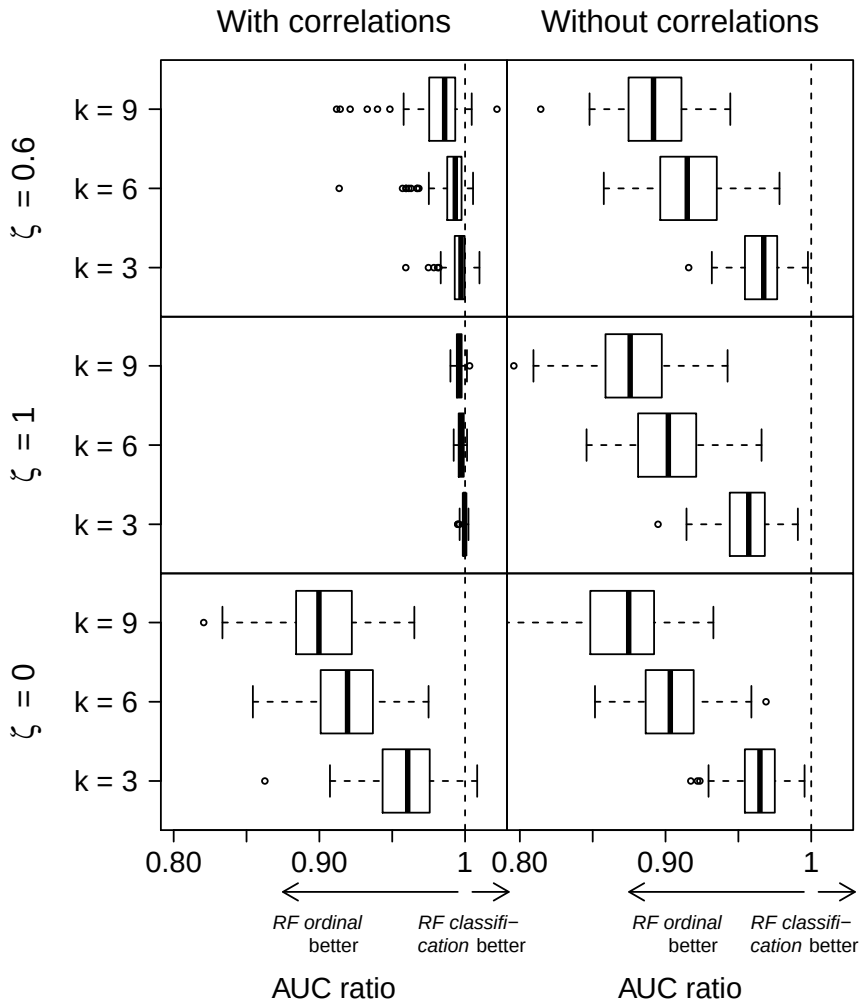

Figure C.4.: Performance ratio for *RF classification* versus *RF ordinal* for simulated data with $p = 1015$ predictor variables for $n = 200$ observations. A ratio of the area under the curve (AUC) below 1 indicates a better prediction accuracy of *RF ordinal* and a ratio above 1 indicates a better prediction accuracy of *RF classification*. Data was generated from a mixture of proportional odds models with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row). Data was generated for $k \in \{3, 6, 9\}$ ordered response levels and for settings in which predictors correlate (left column) and in which all predictors are uncorrelated (right column).

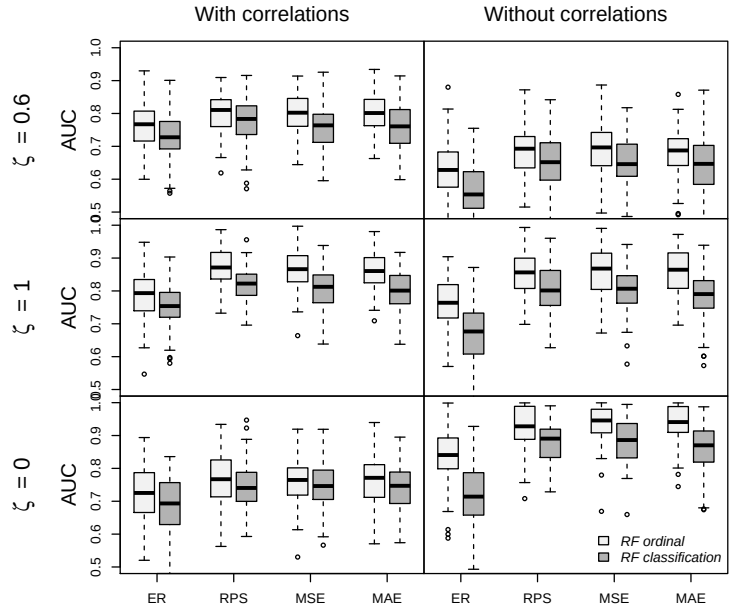

Figure C.5.: Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 9-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

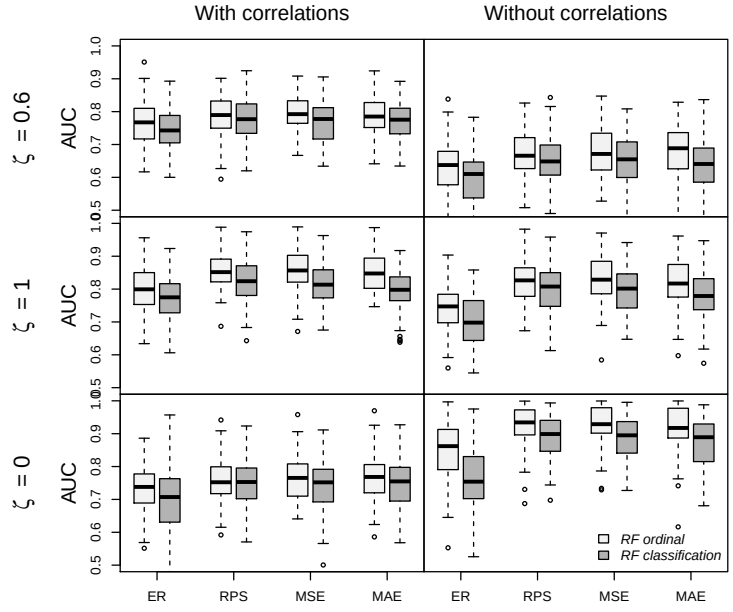

Figure C.6.: Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 6-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

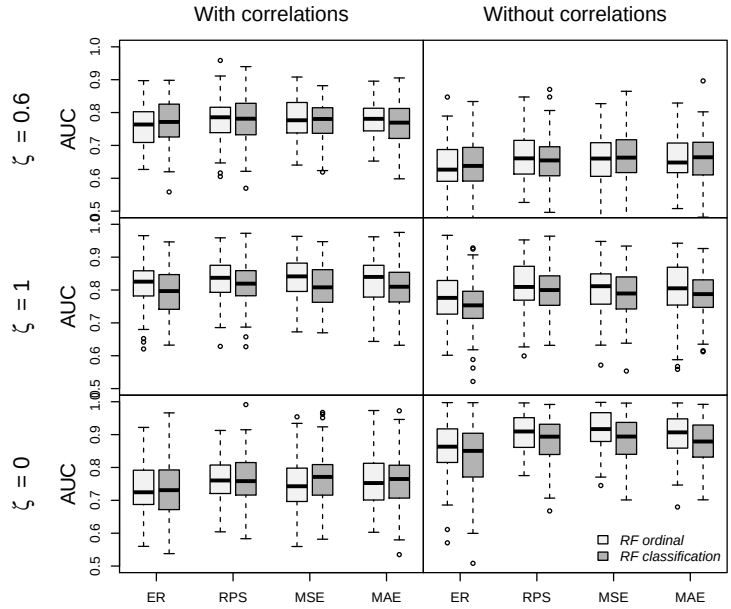

Figure C.7.: Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 3-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

C.2. A variable importance test for high-dimensional data

C.2.1. Studies with complete predictor space

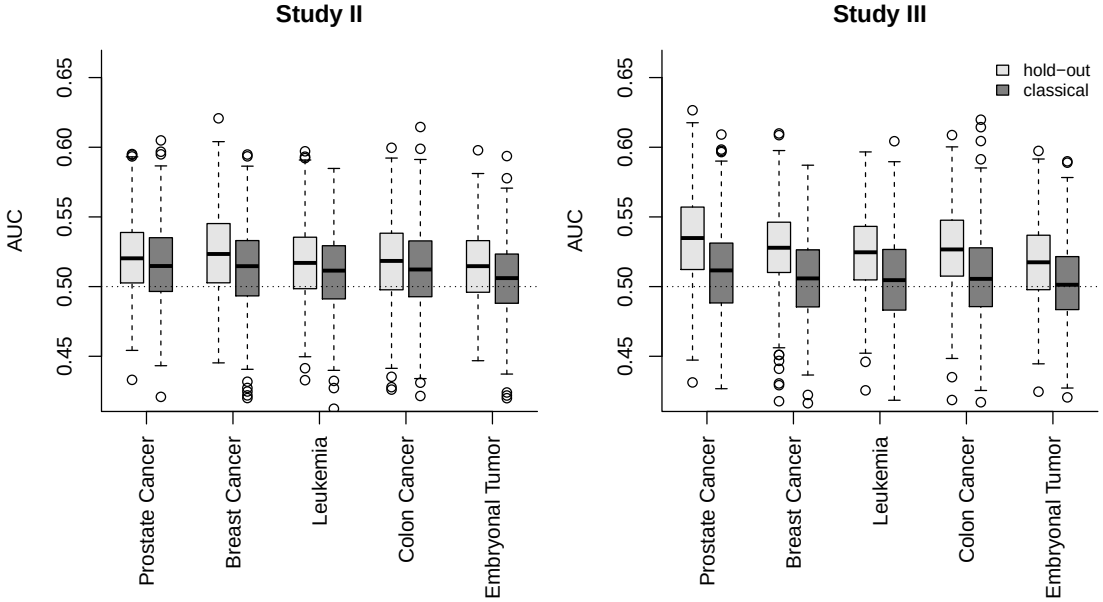

Figure C.8.: Discriminative ability of the novel hold-out VIM and the classical VIM for Study II (left) and Study III (right) with $mtry = \frac{p}{5}$. Discriminative ability is measured by the area under the curve (AUC). Values of 0.5 indicate no discriminative ability (horizontal dotted line).

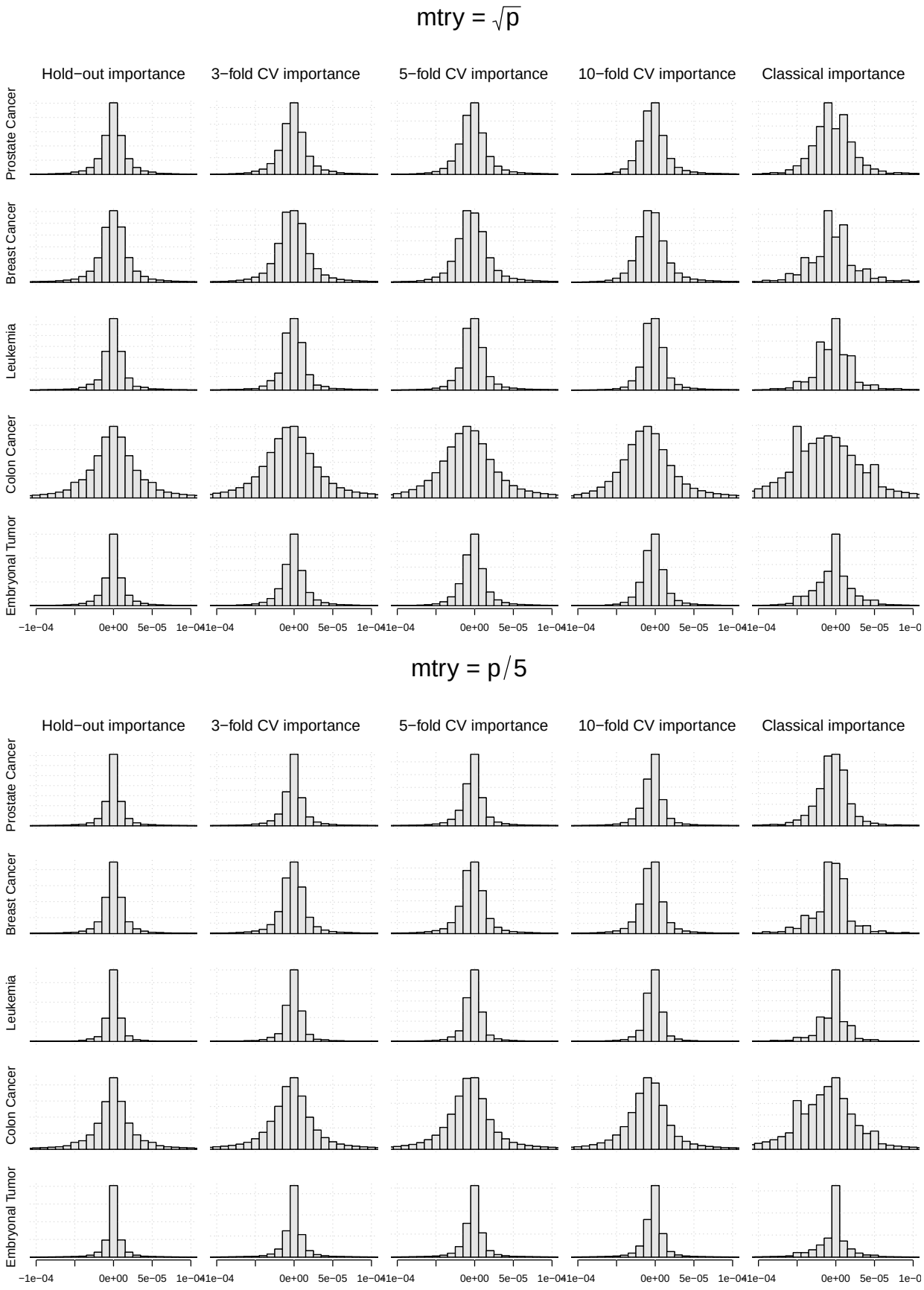

Figure C.9.: Variable importance null distribution when using the hold-out VIM, the cross-validated VIM with $k = 3$, $k = 5$, and $k = 10$ and the classical VIM and setting *mtry* to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower). Distributions are shown for 500 repetitions of Study I.

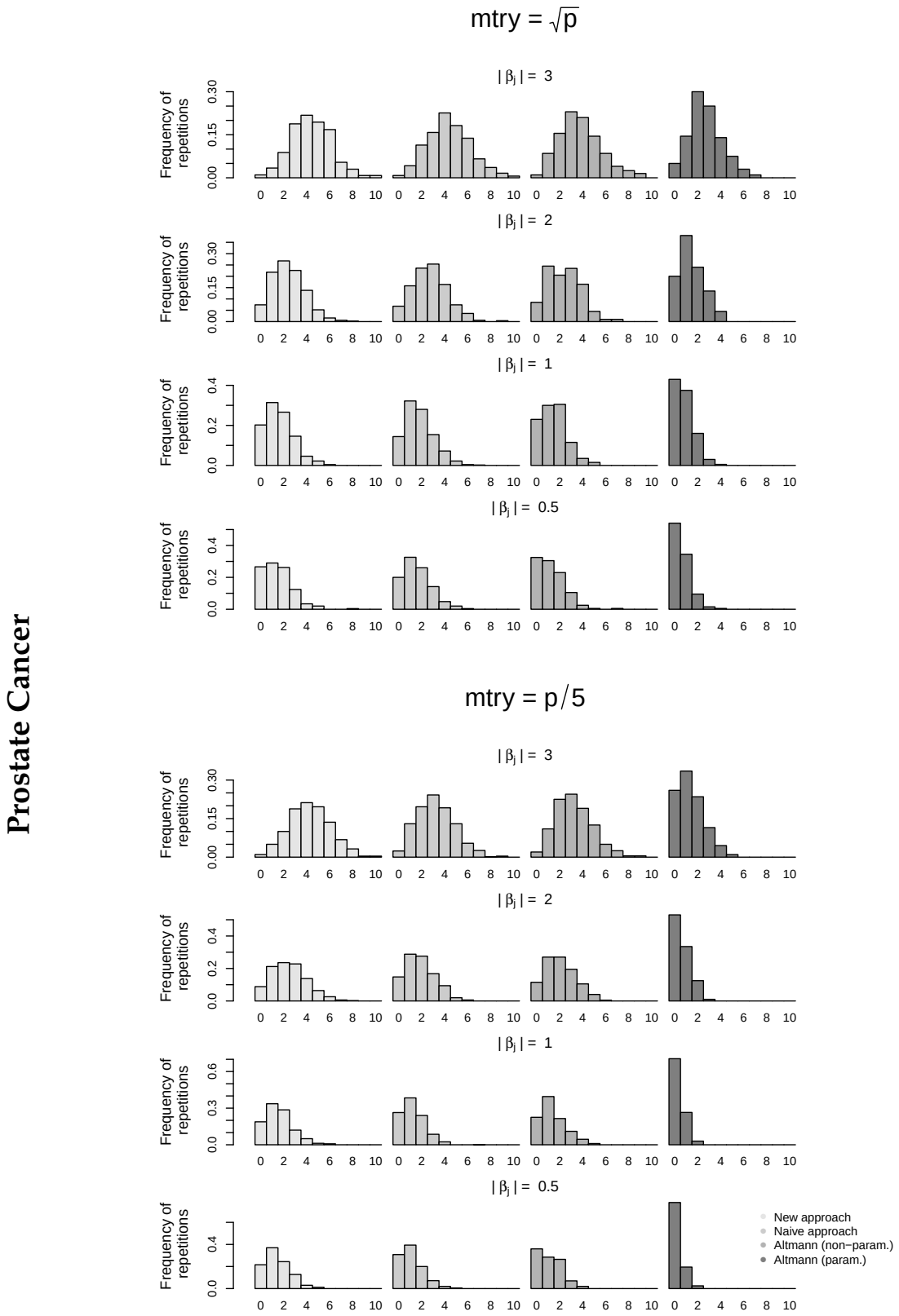

Figure C.10.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

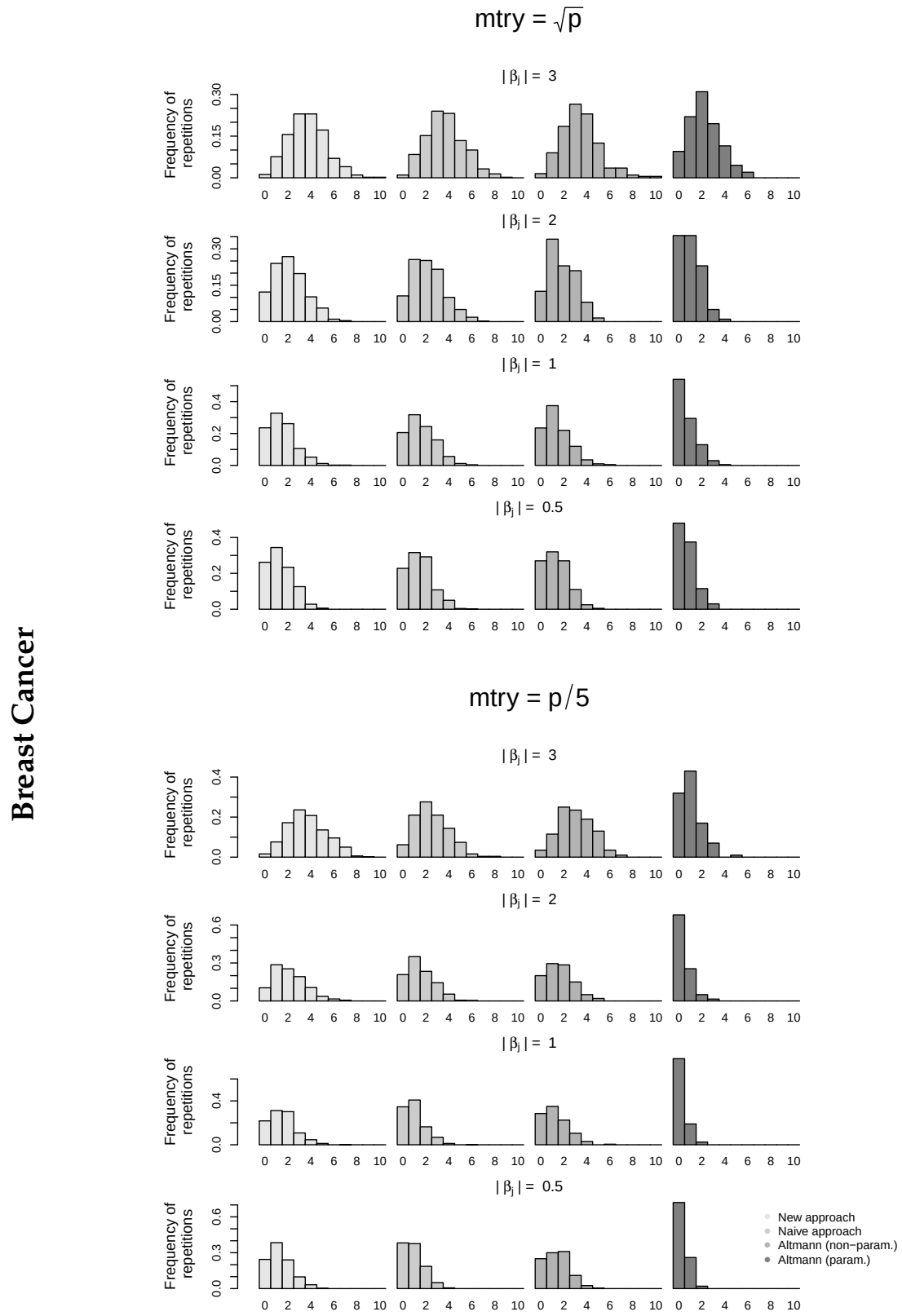

Figure C.11.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

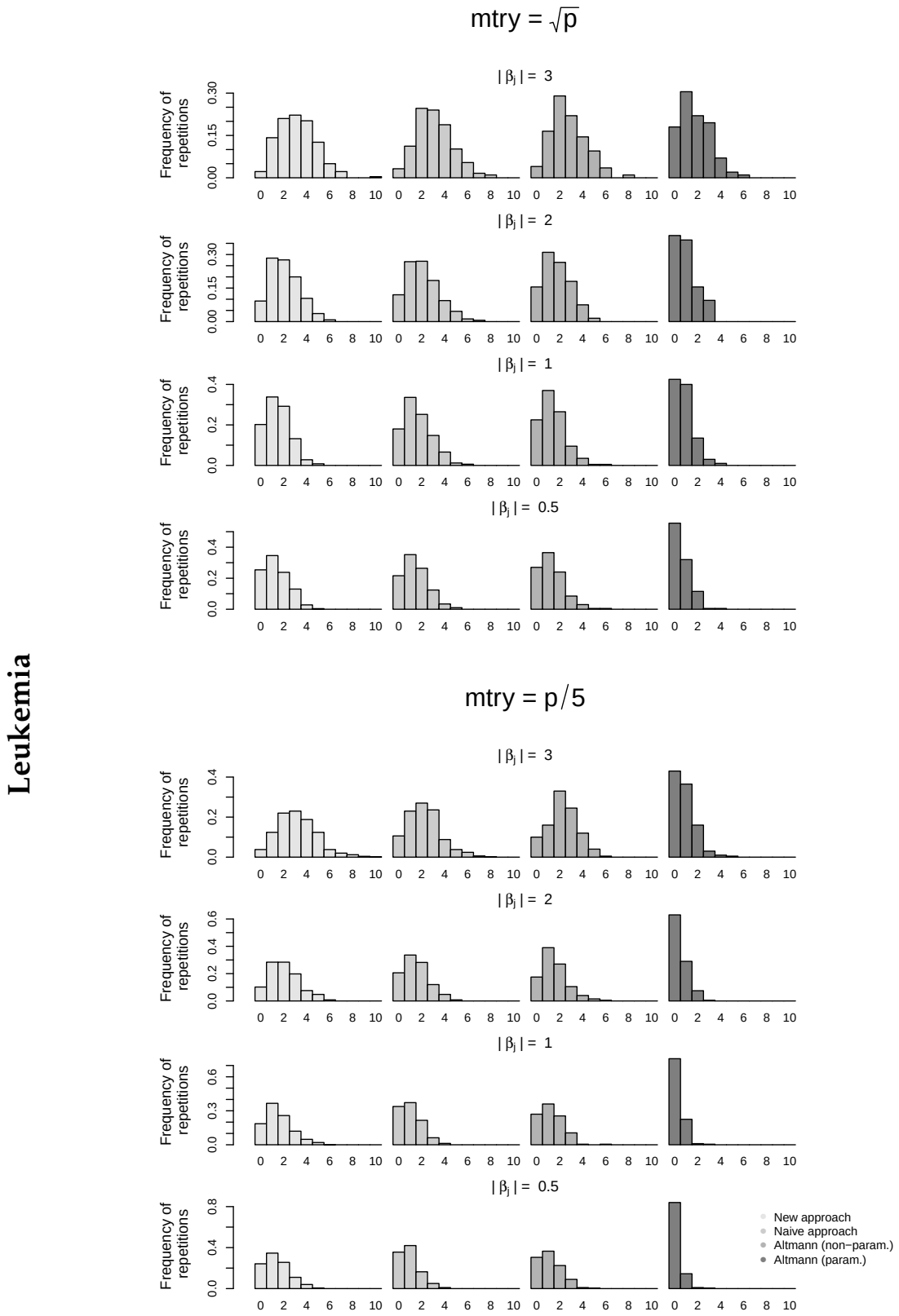

Figure C.12.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

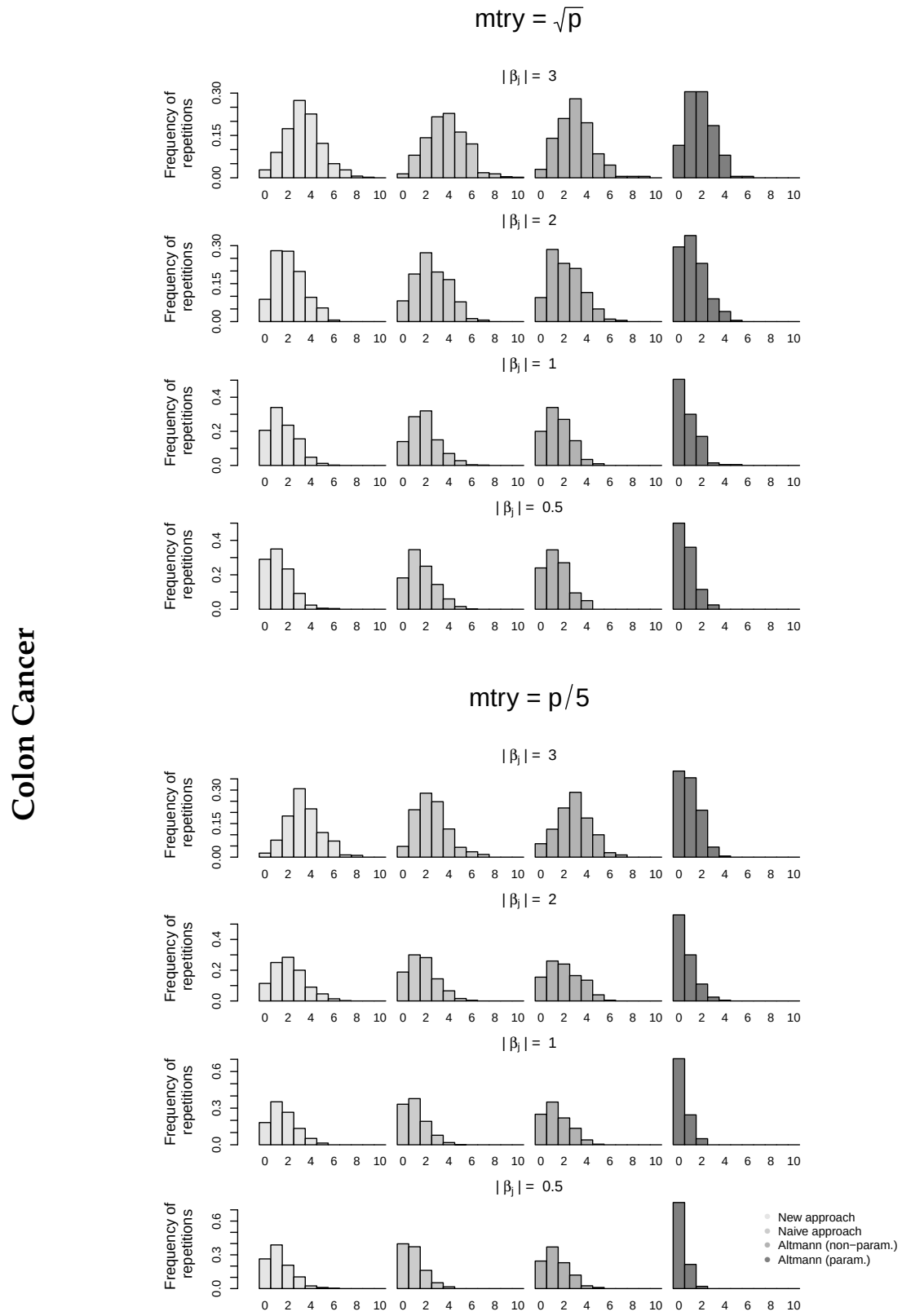

Figure C.13.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

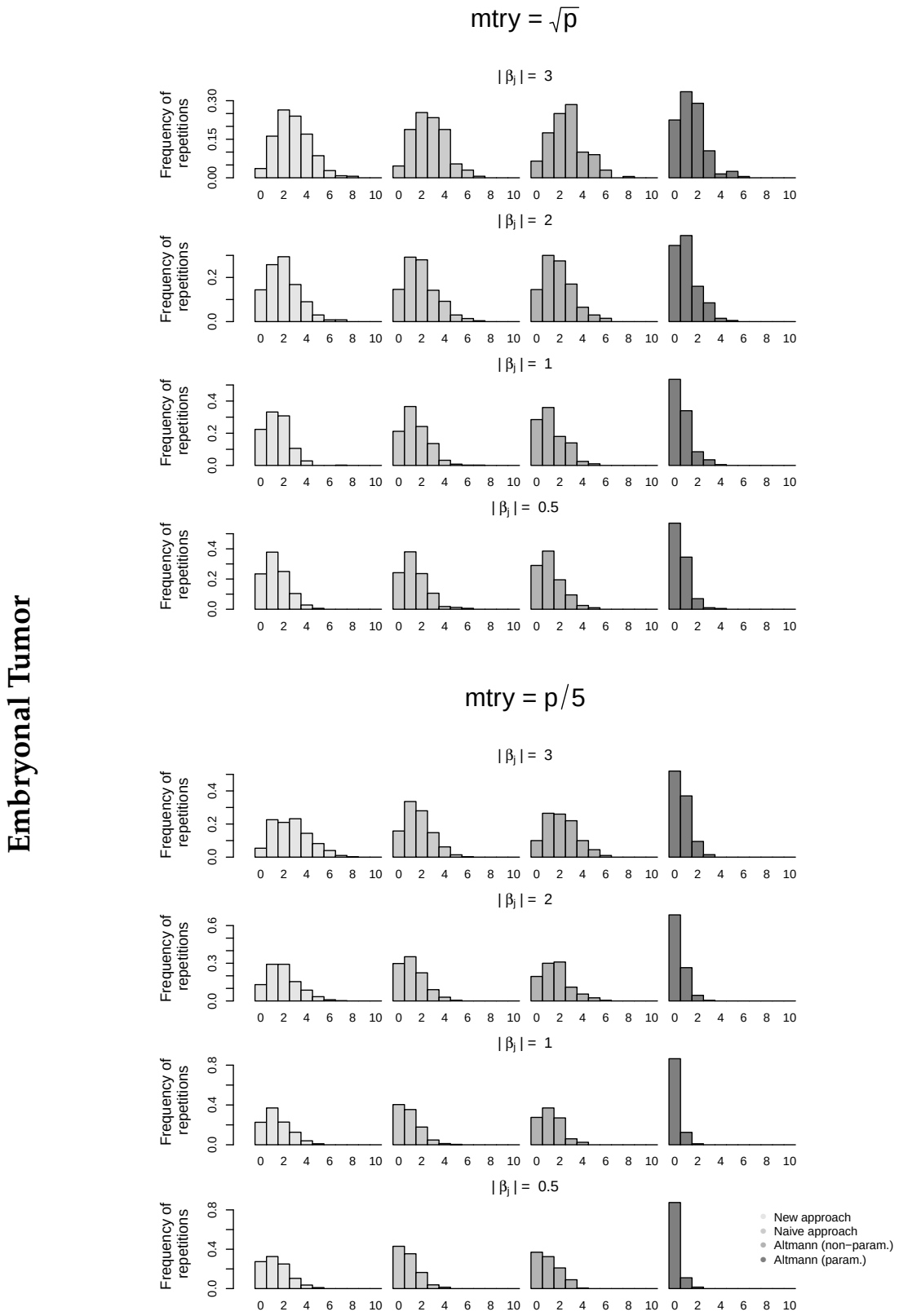

Figure C.14.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

C.2.2. Studies with reduced predictor space $(p = 100)$

Study I

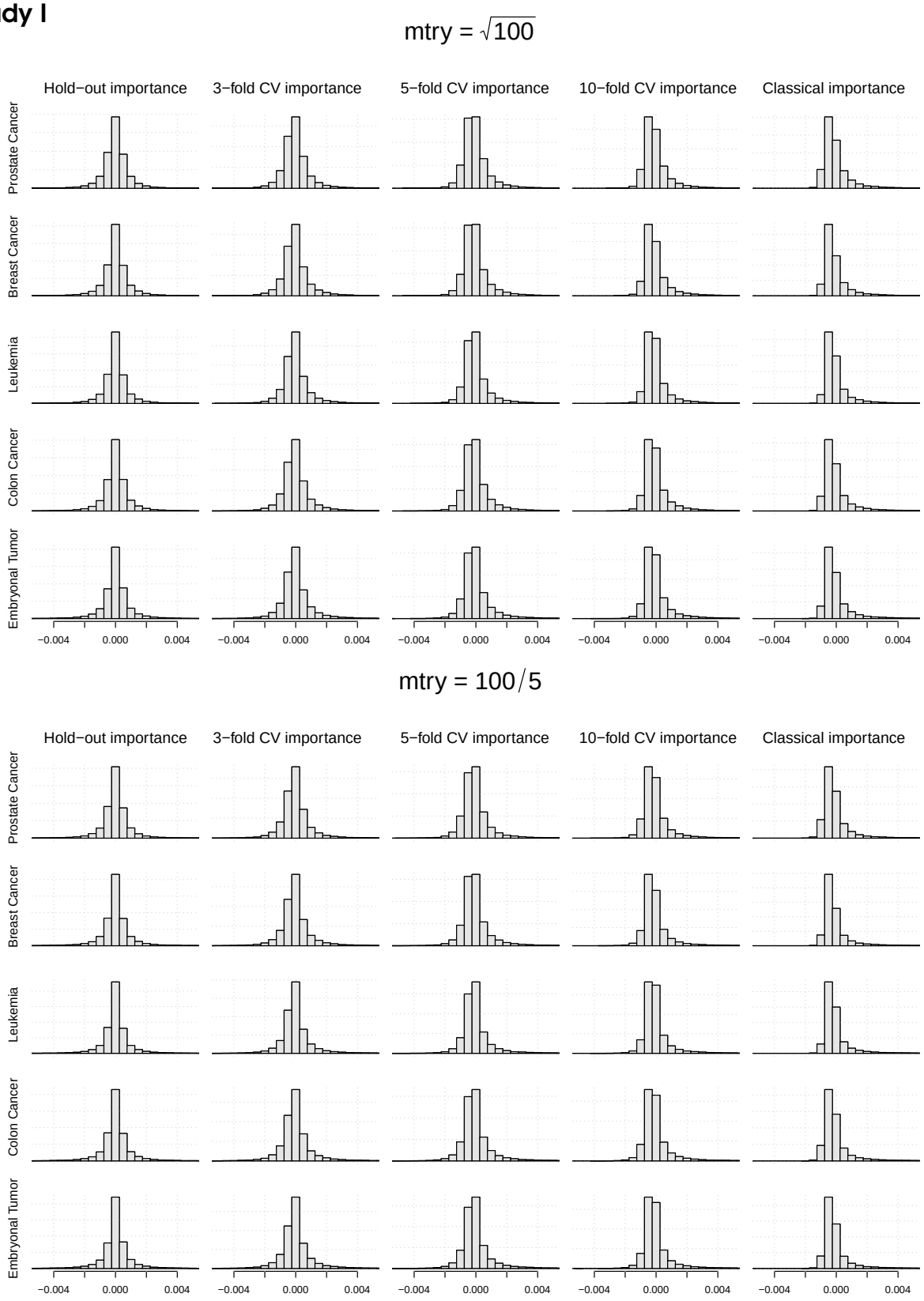

Figure C.15.: Variable importance null distribution when using the hold-out VIM, the cross-validated VIM with $k = 3$, $k = 5$, and $k = 10$ and the classical VIM and setting *mtry* to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower). Distributions are shown for 500 repetitions of Study I.

Study II

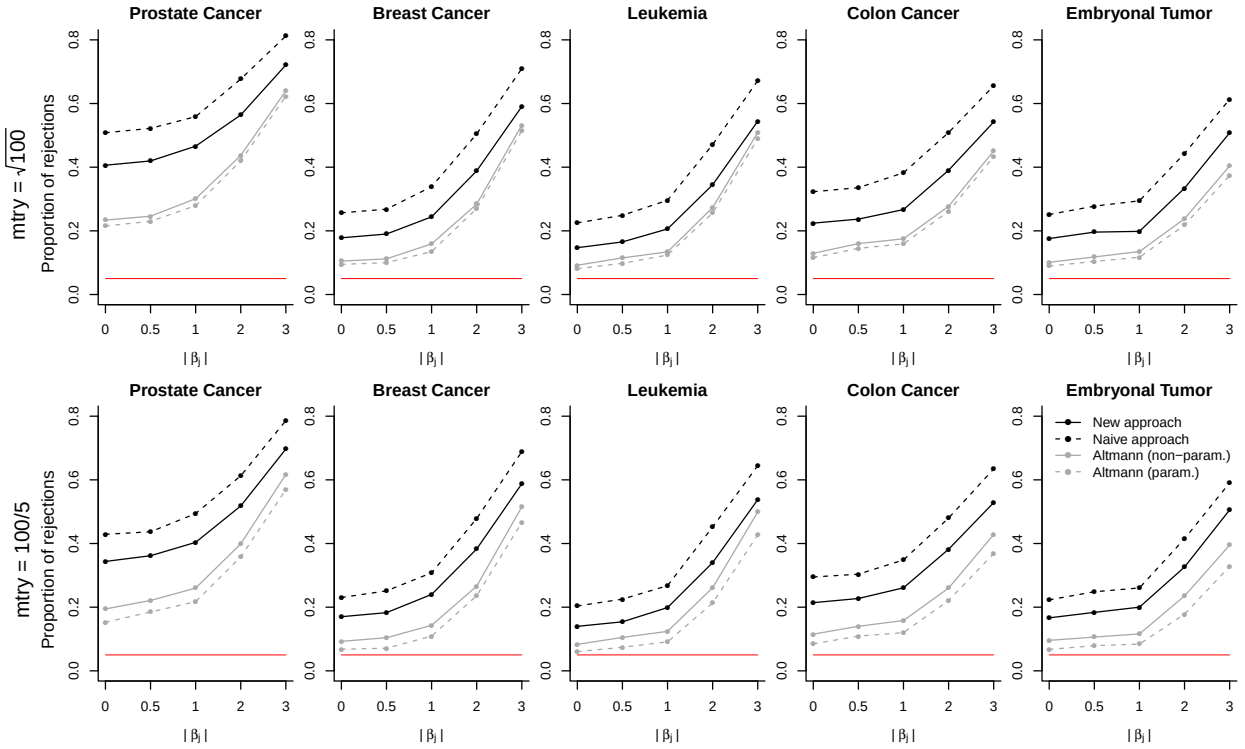

Figure C.16.: Proportion of rejected null hypothesis among predictor variables X_j with specified absolute effect size $|\beta_j| \in \{0, 0.5, 1, 2, 3\}$. The mean proportions over 500 (200 for the approach of Altmann et al. (2010), resp.) repetitions of Study III are shown when using the novel approach, the naive approach and the approach of Altmann et al. (2010), with *mtry* set to $\sqrt{100}$ (upper panel) and $\frac{100}{5}$ (lower panel). The red horizontal line represents the 5% significance level.

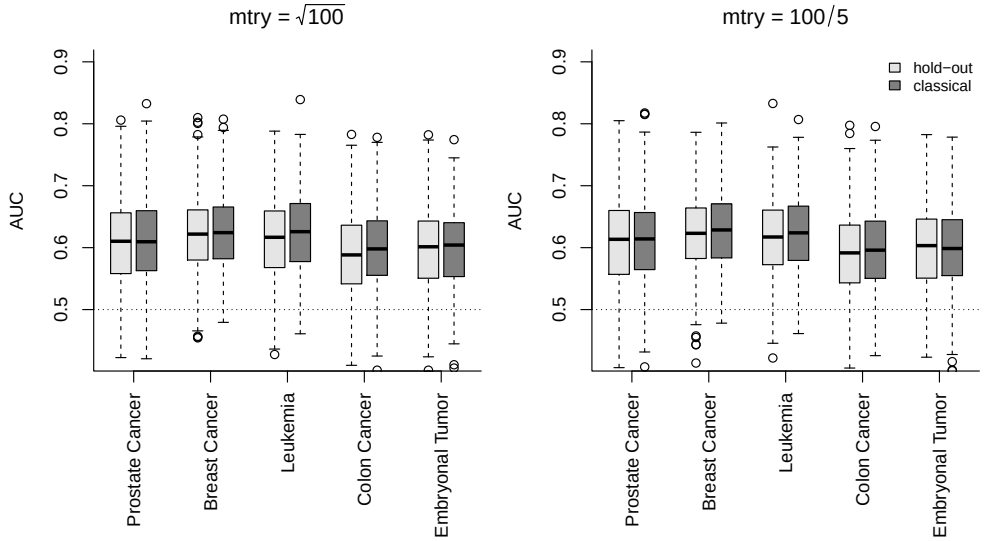

Figure C.17.: Discriminative ability of the novel hold-out VIM and the classical VIM with *mtry* set to $\sqrt{100}$ (left) and $\frac{100}{5}$ (right). Discriminative ability is measured by the area under the curve (AUC). Values of 0.5 indicate no discriminative ability (horizontal dotted line).

Study III

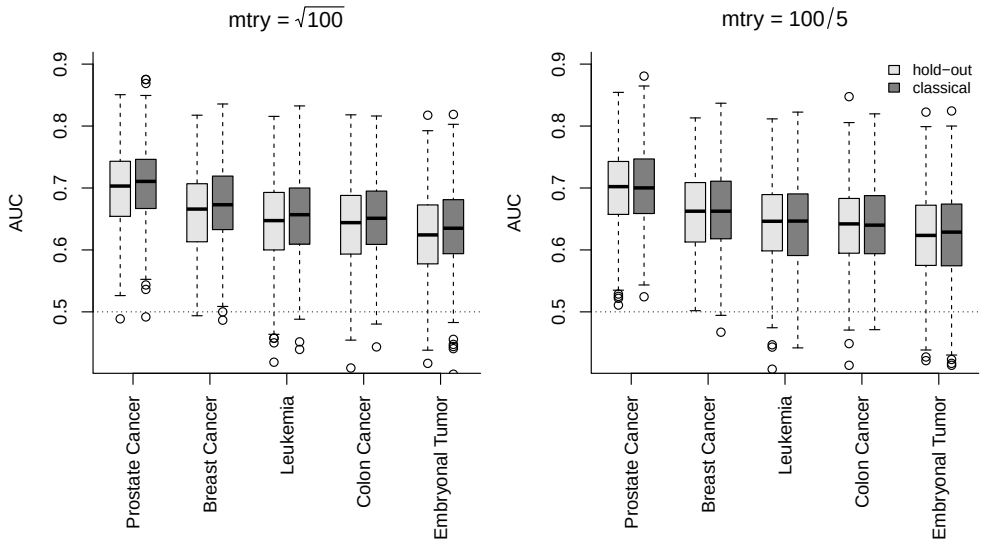

Figure C.18.: Discriminative ability of the novel hold-out VIM and the classical VIM with *mtry* set to $\sqrt{100}$ (left) and $\frac{100}{5}$ (right). Discriminative ability is measured by the area under the curve (AUC). Values of 0.5 indicate no discriminative ability (horizontal dotted line).

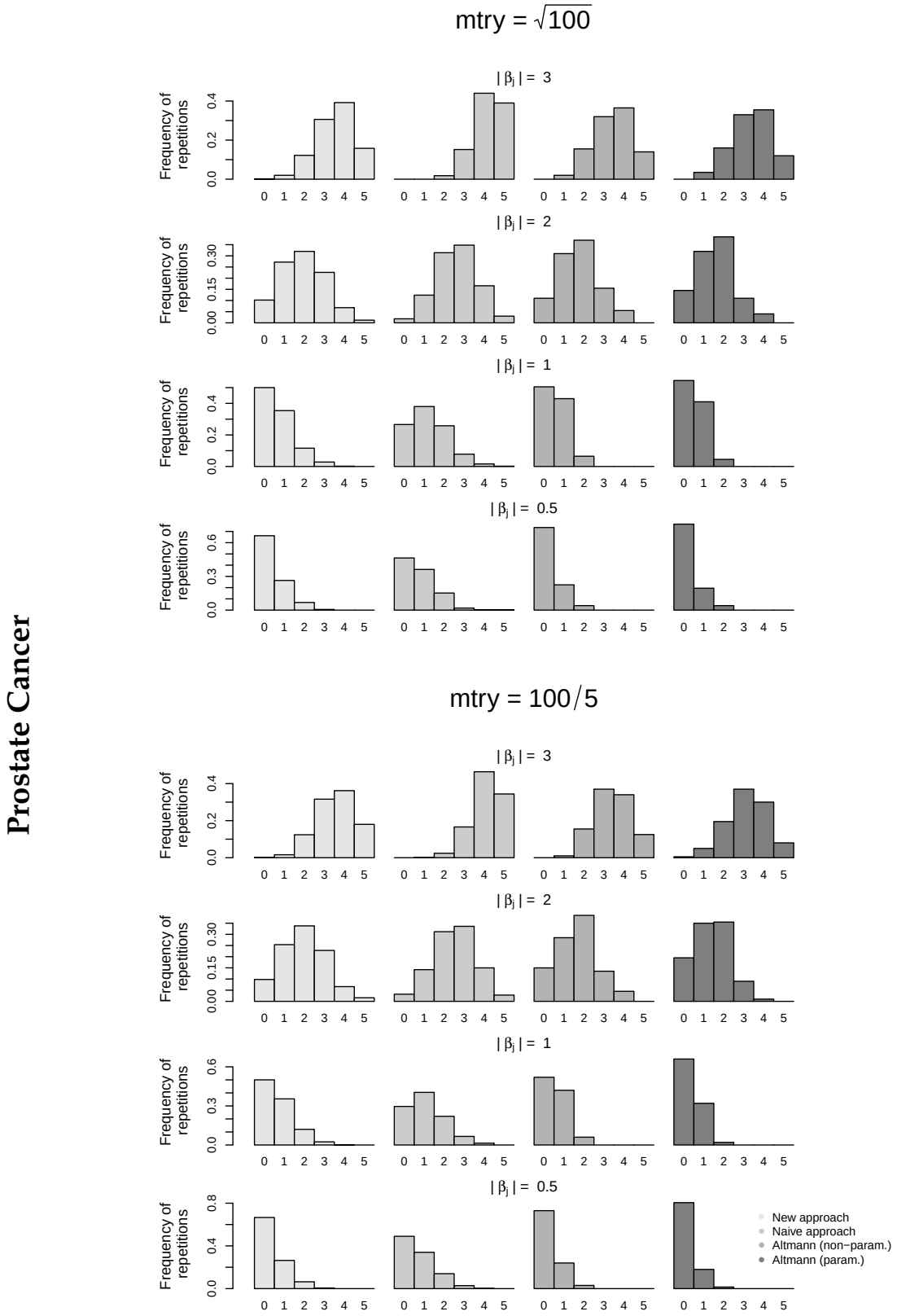

Figure C.19.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

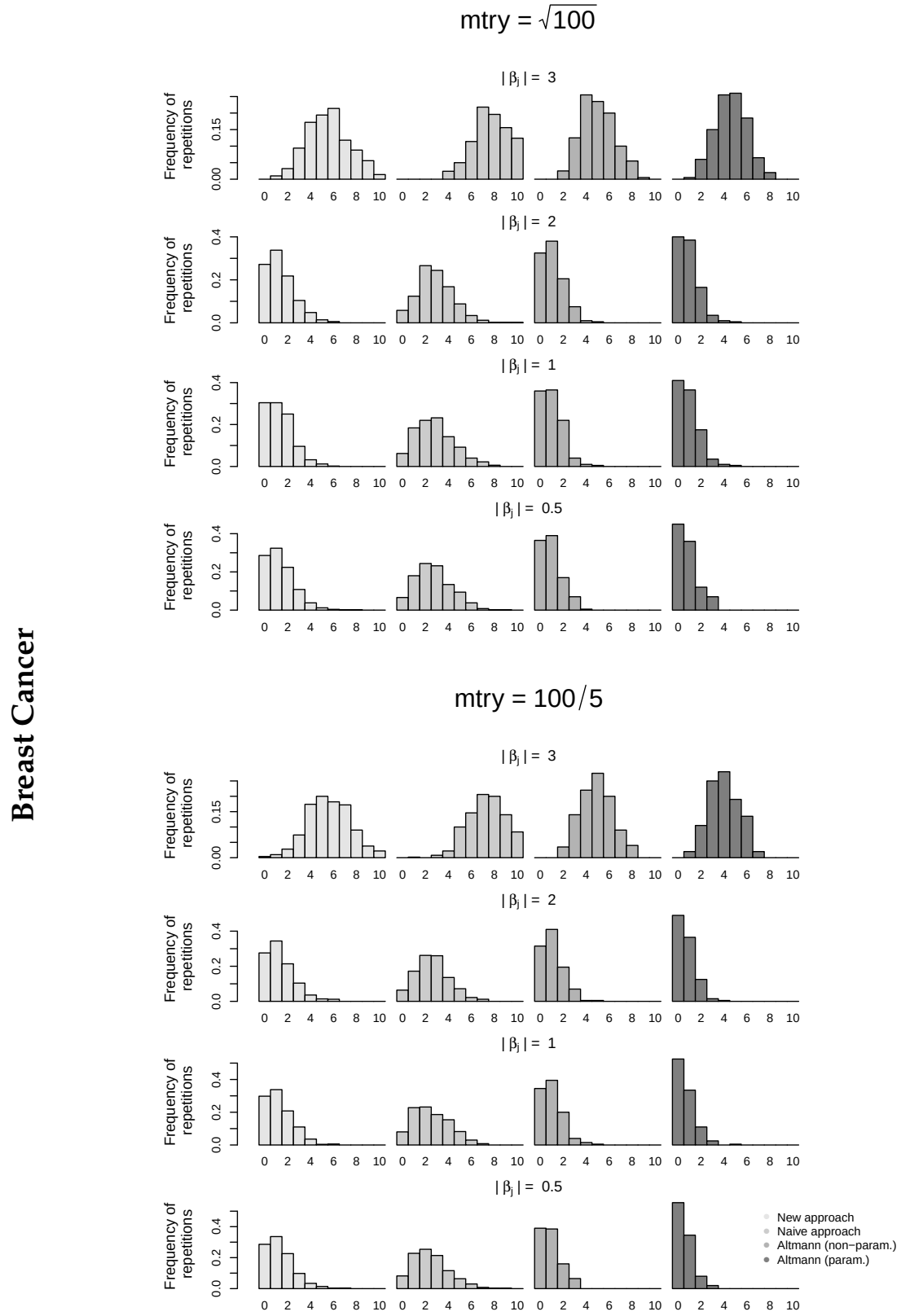

Figure C.20.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

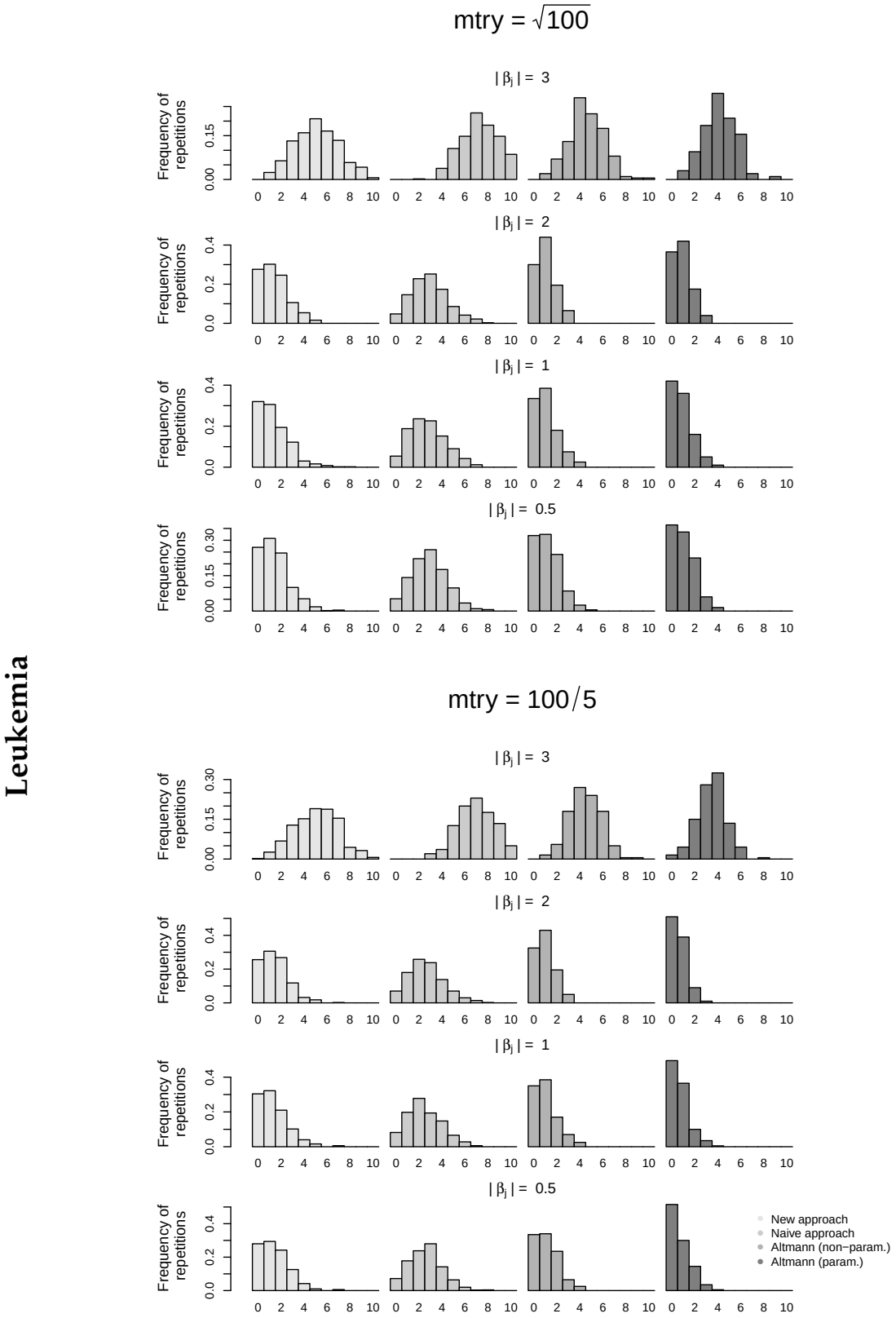

Figure C.21.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

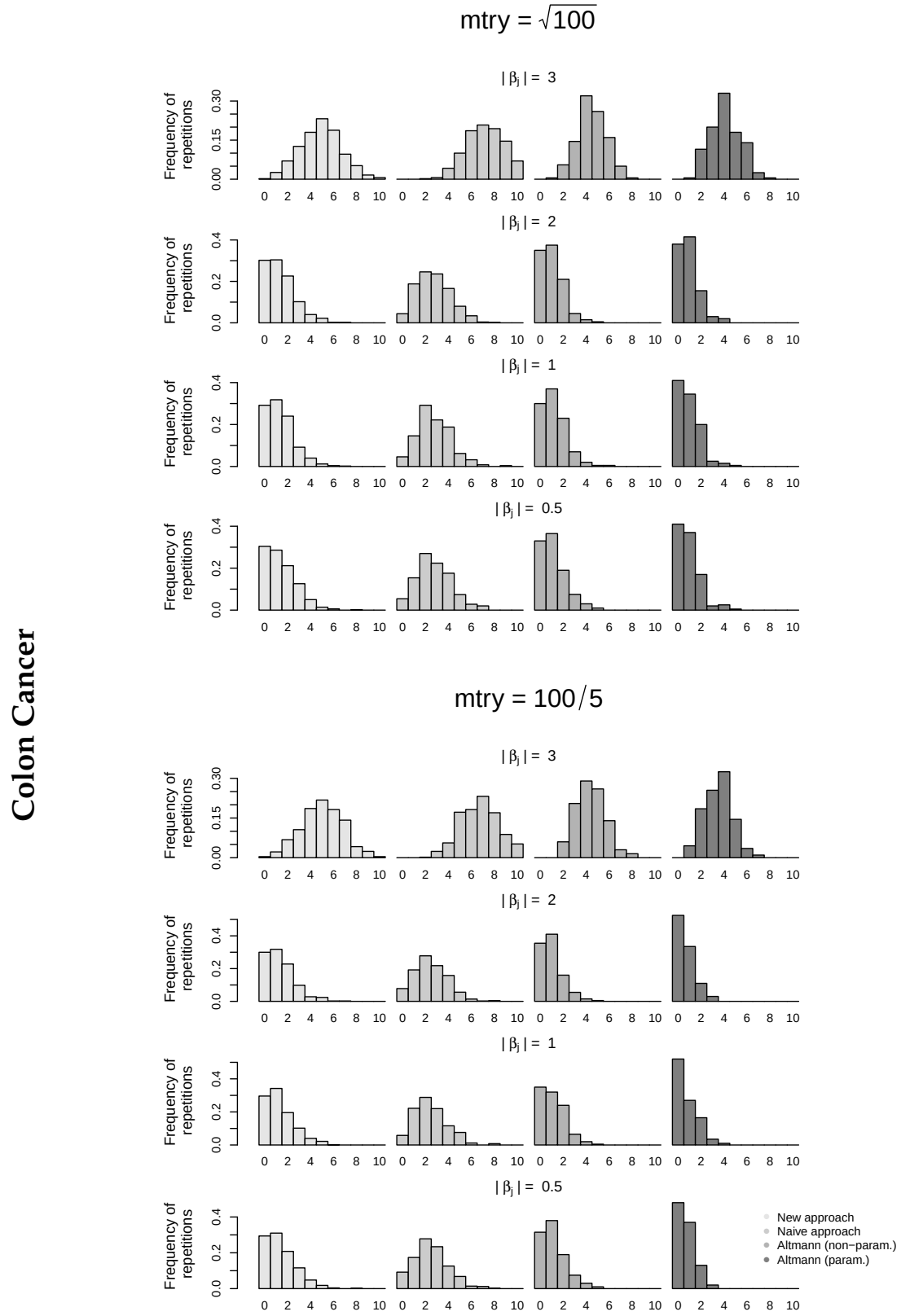

Figure C.22.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

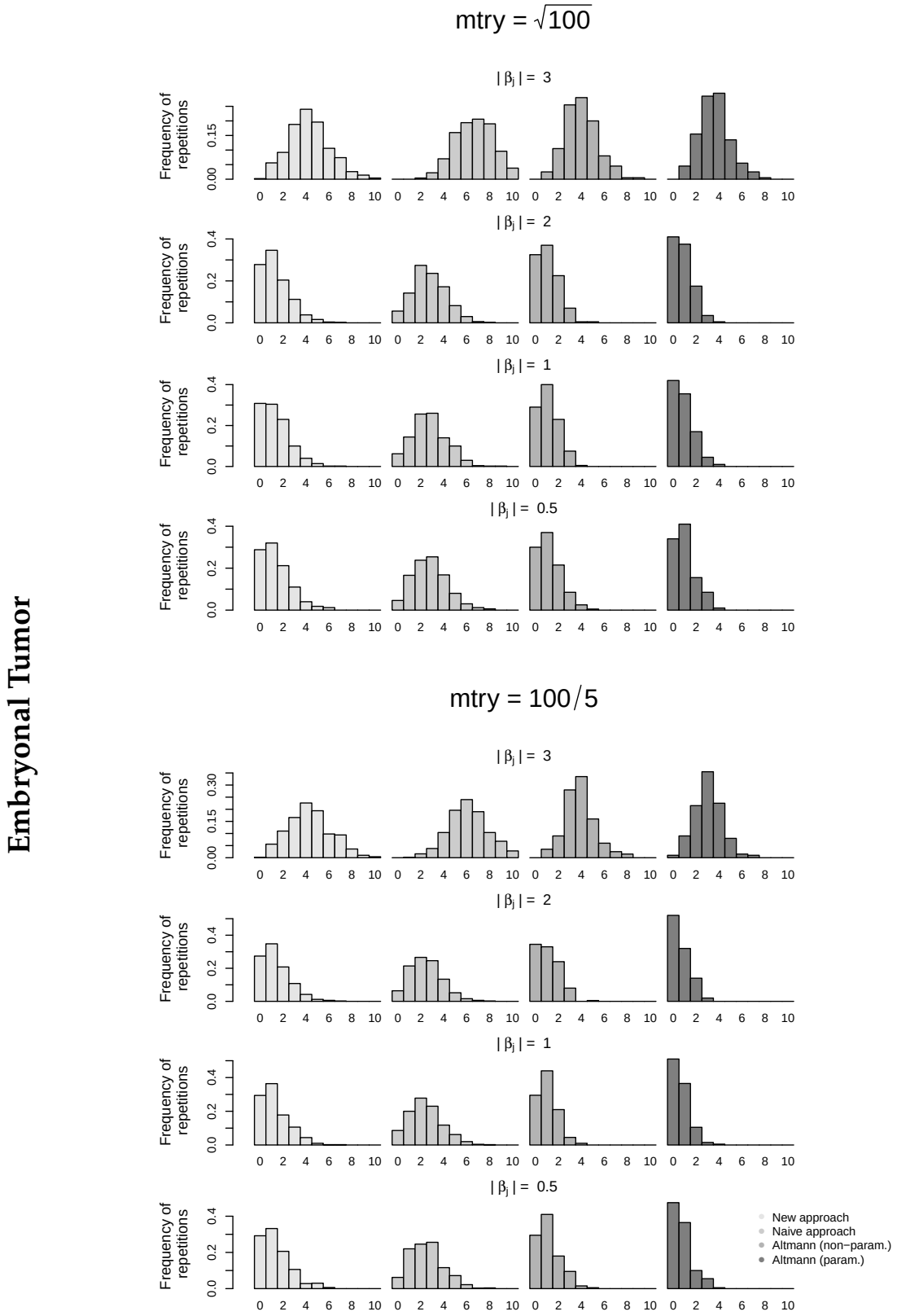

Figure C.23.: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using the new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with *mtry* set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

C.3. Hypothesis tests on bootstrap samples

C.3.1. Empirical studies on the marginal distribution of a bootstrapped Z-test statistic

For computing Z and Z^*, $n = 1000$ independent observations are drawn from the standard normal distribution. Then a bootstrap sample is drawn out of this original sample and the test statistic for a Z-test with null hypothesis $H_0 : \mu = 0$ is computed from both original and bootstrap samples. This procedure is repeated 500000 times, yielding 500000 values of both Z and Z^*. Figure C.24 shows the resulting empirical density functions of Z and Z^*. As expected from theory the distribution of the test statistic Z coincides with the standard normal distribution: the two lines in Figure C.24 completely overlap. The distribution of the test statistic Z^* in contrast systematically deviates from the standard normal distribution. There is a remarkable difference in variances of the test statistics Z and Z^*, while the expected values seem to be equal. The empirical expectation of Z and Z^* are both close to zero with values -0.0010 and 0.0018, respectively. In contrast, the empirical variance of Z^* is, at 2.0011, larger by a factor of 2 than the variance of Z, which is, at 1.0009, very close to the variance of the standard normal distribution.

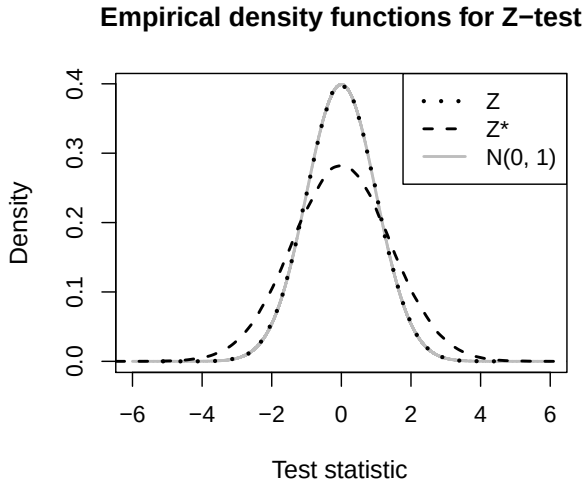

Figure C.24.: Empirical density functions for test statistics Z (dotted black line) and Z^* (dashed black line) of the Z-test. The density of the standard normal distribution is indicated by the solid gray line.

C.3.2. Additional results of the real data application

Scale	Variable	Original rank	Bootstrap rank	(diff.)	Subsample rank	(diff.)
metric or $m = 2$	diabetes	3	5	(-2)	2	$(+1)$
	asthma	4	7	(-3)	4	0
	heartFailure	6	8	(-2)	5	$(+1)$
	AcuteIllness	8	11	(-3)	8	0
	BPsys	10	12	(-2)	10	0
	age	11	13	(-2)	11	0
	alcohol	12	19	(-7)	12	0
	BPdias	13	18	(-5)	13	0
	stroke	14	21	(-7)	14	0
	heavyDrinker	15	22	(-7)	16	(-1)
	sex	16	20	(-4)	15	$(+1)$
	chronicBronchitis	18	17	$(+1)$	18	0
	WBCcount	19	28	(-9)	24	(-5)
	waistcircum	22	23	(-1)	19	$(+3)$
	BMI	23	26	(-3)	23	0
	Cholesterol	25	24	$(+1)$	20	$(+5)$
	100cig	27	25	$(+2)$	22	5
$m = 4$	depression	20	16	$(+4)$	25	(-5)
	country_of_birth	28	27	$(+1)$	28	0
$m = 5$	sleepTrouble	1	2	(-1)	1	0
	medicalPlaceToGo	5	4	$(+1)$	6	(-1)
	wakeUp	9	6	$(+3)$	9	0
	race	17	9	$(+8)$	17	0
	education	21	10	$(+11)$	21	0
	HealthStatus	24	14	$(+10)$	26	(-2)
	ToothCond	26	15	$(+11)$	27	(-1)
$m = 6$	marital_status	7	3	$(+4)$	7	0
$m = 12$	income	2	1	$(+1)$	3	(-1)

Table C.1.: Variable ranking for the first of the 1000 permuted NHANES data sets (modification consisted of permuting the response variable). Variable rankings are determined by p-values obtained for the original sample ("Original rank"), by the median bootstrapped p-value ("Bootstrap rank"), and by the median p-value from subsamples ("Subsample rank"). The difference to the "Original rank" is given in brackets for each variable. The parameter m denotes the number of levels for the categorical predictor variables.

Bibliography

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators, *Econometrica* **76**(6): 1537–1557.

Agresti, A. (2002). *Categorical data analysis*, Vol. 359, John Wiley & Sons, Hoboken, New Jersey.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *in* B. Petrov and F. Caski (eds), *Second International Symposium on Information Theory*, Akademia Kiado, Budapest, pp. 267–281.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences* **96**(12): 6745–6750.

Altman, D. G. and Andersen, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model, *Statistics in Medicine* **8**(7): 771–783.

Altmann, A., Toloşi, L., Sander, O. and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure, *Bioinformatics* **26**(10): 1340–1347.

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space, *Econometrica* **68**(2): 399–405.

Archer, K. and Mas, V. (2009). Ordinal response prediction using bootstrap aggregation, with application to a high-throughput methylation data set, *Statistics in Medicine* **28**(29): 3597–3610.

Bath, P., Geeganage, C., Gray, L., Collier, T. and Pocock, S. (2008). Use of ordinal outcomes in vascular prevention trials: Comparison with binary outcomes in published trials, *Stroke* **39**(10): 2817–2823.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, *Machine Learning* **36**(1): 105–139.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap, *The Annals of Statistics* **9**(6): 1196–1217.

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

Black, S., Kushner, I. and Samols, D. (2004). C-reactive protein, *Journal of Biological Chemistry* **279**(47): 48487–48490.

Bollen, K. A. and Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models, *Sociological Methods & Research* **21**(2): 205–229.

Boos, D. D. and Stefanski, L. A. (2011). P-value precision and reproducibility, *The American Statistician* **65**(4): 213–221.

Boulesteix, A.-L. (2015). Ten simple rules for reducing overoptimistic reporting in methodological computational research, *PLoS Computational Biology* **11**(4): e1004191.

Boulesteix, A.-L., Bender, A., Bermejo, L. J. and Strobl, C. (2012). Random forest Gini importance favours SNPs with large minor allele frequency: assessment, sources and recommendations, *Briefings in Bioinformatics* **13**(3): 292–304.

Boulesteix, A.-L., Janitza, S., Hapfelmeier, A., Van Steen, K. and Strobl, C. (2015). Letter to the Editor: On the term 'interaction' and related phrases in the literature on random forests, *Briefings in Bioinformatics* **16**(2): 338–345.

Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**(6): 493–507.

Boulesteix, A.-L. and Strobl, C. (2009). Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction, *BMC Medical Research Methodology* **9**: 85.

Boulesteix, A.-L., Strobl, C., Augustin, T. and Daumer, M. (2008). Evaluating microarray-based classifiers: an overview, *Cancer Informatics* **6**: 77–97.

Braga-Neto, U. M. and Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification?, *Bioinformatics* **20**(3): 374–380.

Breiman, L. (1996a). Bagging predictors, *Machine Learning* **24**(2): 123–140.

Breiman, L. (1996b). Out-of-bag estimation, *Technical report*, Department of Statistics, University of California.
URL: *https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf*

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

Breiman, L. and Cutler, A. (2004). Random forests, `http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm`. Accessed: 2015-12-30.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and regression trees*, Chapman & Hall, Boca Raton, Florida.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting, *Statistical Science* **22**(4): 477–505.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multi-model inference: a practical information-theoretic approach*, Springer, New York.

Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates, *Machine Learning* **48**(1): 287–297.

Celik, E. (2015). *Erweiterung eines Testansatzes für die Variablenwichtigkeit in Random Forests*, Master's thesis, University of Munich, Germany.

Chen, C.-H. and George, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model, *Statistics in Medicine* **4**(1): 39–46.

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers*, John Wiley & Sons, Hoboken, New Jersey.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties, *Decision Support Systems* **47**(4): 547–553.

Daumer, M., Held, U., Ickstadt, K., Heinz, M., Schach, S. and Ebers, G. (2008). Reducing the probability of false positive research findings by pre-publication validation – experience with a large multiple sclerosis database, *BMC Medical Research Methodology* **8**: 18.

Davison, A. C. (1997). *Bootstrap methods and their application*, Cambridge University Press, Cambridge.

Davison, A. C., Hinkley, D. V. and Young, G. A. (2003). Recent developments in bootstrap methodology, *Statistical Science* **18**(2): 141–157.

De Bin, R., Janitza, S., Sauerbrei, W. and Boulesteix, A.-L. (2016). Subsampling versus bootstrap in resampling-based model selection for multivariable regression, *Biometrics* **72**(1): 272–280.

Dettling, M. and Bühlmann, P. (2003). Boosting for tumor classification with gene expression data, *Bioinformatics* **19**(9): 1061–1069.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest, *BMC Bioinformatics* **7**: 3.

Dietterich, T. (2000). Ensemble methods in machine learning, *Multiple Classifier Systems*, Vol. 1857 of *Lecture Notes in Computer Science*, Springer, pp. 1–15.

Dolch, M. E., Janitza, S., Boulesteix, A.-L., Grassmann, C., Praun, S., Denzer, W., Schelling, G. and Schubert, S. (2016). Gram-negative and -positive bacteria differentiation in blood culture samples by headspace volatile compound analysis, *Journal of Biological Research-Thessaloniki* **23**: 3.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*, Chapman & Hall/CRC, Boca Raton, Florida.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories, *Journal of Applied Meteorology* **8**(6): 985–987.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics* **29**(5): 1189–1232.

Fürnkranz, J. and Hüllermeier, E. (2010). *Preference learning*, Springer, Berlin.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**(5439): 531–537.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses*, Springer, New York.

Hapfelmeier, A. and Ulm, K. (2013). A new variable selection approach using random forests, *Computational Statistics & Data Analysis* **60**: 50–69.

Harrington, D. L., Liu, D., Smith, M. M., Mills, J. A., Long, J. D., Aylward, E. H. and Paulsen, J. S. (2014). Neuroanatomical correlates of cognitive functioning in prodromal Huntington disease, *Brain and Behavior* **4**(1): 29–40.

Hartigan, J. A. (1969). Using subsample values as typical values, *Journal of the American Statistical Association* **64**(328): 1303–1317.

Hastie, T. (2007). Comment: Boosting algorithms: Regularization, prediction and model fitting, *Statistical Science* **22**(4): 513–515.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning*, Springer Series in Statistics, Springer, New York.

Hechenbichler, K. and Schliep, K. (2004). Weighted k-nearest-neighbor techniques and ordinal classification, Discussion Paper 399, University of Munich.
URL: *https://epub.ub.uni-muenchen.de/1769/1/paper_399.pdf*

Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost, *Computational Statistics* **29**(1): 3–35.

Hosmer Jr, D. W. and Lemeshow, S. (2004). *Applied logistic regression*, John Wiley & Sons, New York.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2010). Model-based boosting 2.0, *The Journal of Machine Learning Research* **11**: 2109–2113.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2013). *mboost: Model-Based Boosting*. R package version 2.2-3.
URL: *http://CRAN.R-project.org/package=mboost*

Hothorn, T., Hornik, K., Van De Wiel, M. A. and Zeileis, A. (2006). A lego system for conditional inference, *The American Statistician* **60**(3): 257–263.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics* **15**(3): 651–674.

Hothorn, T., Hornik, K. and Zeileis, A. (2012). Party: a laboratory for recursive partytioning, *R package version 1.0-3* .
URL: *http://cran.r-project.org/package=party*

Hothorn, T., Leisch, F., Zeileis, A. and Hornik, K. (2005). The design and analysis of benchmark experiments, *Journal of Computational and Graphical Statistics* **14**(3): 675–699.

Huynh-Thu, V. A., Saeys, Y., Wehenkel, L. and Geurts, P. (2012). Statistical interpretation of machine learning-based feature importance scores for biomarker discovery, *Bioinformatics* **28**(13): 1766–1774.

Janitza, S., Binder, H. and Boulesteix, A.-L. (2014). Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications, *Technical Report 163*, Department of Statistics, University of Munich. (version of 27 June 2014).

Janitza, S., Binder, H. and Boulesteix, A.-L. (2016). Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications, *Biometrical Journal* **58**(3): 447–473.

Janitza, S., Celik, E. and Boulesteix, A.-L. (2015). A computationally fast variable importance test for random forests for high-dimensional data, *Technical Report 185*, Department of Statistics, University of Munich.

Janitza, S., Strobl, C. and Boulesteix, A.-L. (2013). An AUC-based permutation variable importance measure for random forests, *BMC Bioinformatics* **14**: 119.

Janitza, S., Tutz, G. and Boulesteix, A.-L. (2016). Random forest for ordinal responses: Prediction and variable selection, *Computational Statistics & Data Analysis* **96**: 57–73.

Karamanian, V. A., Harhay, M., Grant, G. R., Palevsky, H. I., Grizzle, W. E., Zamanian, R. T., Ihida-Stansbury, K., Taichman, D. B., Kawut, S. M. and Jones, P. L. (2014). Erythropoietin upregulation in pulmonary arterial hypertension, *Pulmonary Circulation* **4**(2): 269–279.

Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits, *Journal of the American Statistical Association* **96**(454): 589–604.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Dawson, N. V., Fulkerson, W. J., Califf, R. M., Desbiens, N. et al. (1995). The support prognostic model: objective estimates of survival for seriously ill hospitalized adults, *Annals of Internal Medicine* **122**(3): 191–203.

Kohavi, R. and Kunz, C. (1997). Option decision trees with majority votes, *Proceedings of the Fourteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 161–169.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest, *R News* **2**(3): 18–22.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Loh, W.-Y. (2014). Fifty years of classification and regression trees, *International Statistical Review* **82**(3): 329–348.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees, *Statistica Sinica* **7**(4): 815–840.

Louppe, G. (2014). Understanding random forests: From theory to practice, *arXiv preprint arXiv:1407.7502* .

Maclin, R. and Opitz, D. (1997). An empirical evaluation of bagging and boosting, *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, AAAI Press, pp. 546–551.

Manly, B. F. (2006). *Randomization, bootstrap and Monte Carlo methods in biology*, 3 edn, Chapman & Hall/CRC, Boca Raton, Florida.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure, *Journal of the American Statistical Association* **58**(303): 690–700.

Mayr, A., Hofner, B., Schmid, M. et al. (2012). The importance of knowing when to stop, *Methods of Information in Medicine* **51**(2): 178–186.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4): 417–473.

Mitchell, M. W. (2011). Bias of the random forest out-of-bag (OOB) error for certain input parameters, *Open Journal of Statistics* **1**(3): 205–211.

Molinaro, A. M., Carriero, N., Bjornson, R., Hartge, P., Rothman, N. and Chatterjee, N. (2011). Power of data mining methods to detect genetic associations and interactions, *Human Heredity* **72**(2): 85–97.

Mukherjee, S., Roberts, S. J., Sykacek, P. and Gurr, S. J. (2003). Gene ranking using bootstrapped p-values, *ACM SIGKDD Explorations Newsletter* **5**(2): 16–22.

Murphy, A. H. (1970). The ranked probability score and the probability score: A comparison, *Monthly Weather Review* **98**(12): 917–924.

National Center for Health Statistics (2012). NHANES 2007 to 2008 public data general release file documentation, `http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/generaldoc_e.htm`.

Nicodemus, K. (2011). Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures, *Briefings in Bioinformatics* **12**(4): 369–373.

Nicodemus, K. K. and Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies, *Bioinformatics* **25**(15): 1884–1890.

Nicodemus, K., Malley, J., Strobl, C. and Ziegler, A. (2010). The behavior of random forest permutation-based variable importance measures under predictor correlation, *BMC Bioinformatics* **11**: 110.

O'Shea, T. M., Kothadia, J. M., Roberts, D. D. and Dillard, R. G. (1998). Perinatal events and the risk of intraparenchymal echodensity in very-low-birthweight neonates, *Paediatric and Perinatal Epidemiology* **12**(4): 408–421.

Pepe, M. (2004). *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, New York.

Piccarreta, R. (2001). A new measure of nominal-ordinal association, *Journal of Applied Statistics* **28**(1): 107–120.

Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M. S., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J. A. et al. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer, *Nature* **518**(7539): 360–364.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions, *The Annals of Statistics* **22**(4): 2031–2050.

Politis, D., Romano, J. and Wolf, M. (1999). *Subsampling*, Springer Series in Statistics, Springer, New York.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C. et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature* **415**(6870): 436–442.

Prosperi, M. C., Marinho, S., Simpson, A., Custovic, A. and Buchan, I. E. (2014). Predicting phenotypes of asthma and eczema with machine learning, *BMC Medical Genomics* **7**(Suppl 1): S7.

Quinlan, J. R. (1986). Induction of decision trees, *Machine Learning* **1**(1): 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, California.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Reif, D. M., Motsinger-Reif, A. A., McKinney, B. A., Rock, M. T., Crowe, J. and Moore, J. H. (2009). Integrated analysis of genetic and proteomic data identifies biomarkers associated with adverse events following smallpox vaccination, *Genes and Immunity* **10**(2): 112–119.

Rospleszcz, S., Janitza, S. and Boulesteix, A.-L. (2016). Categorical variables with many categories are preferentially selected in model selection procedures for multivariable regression models on bootstrap samples, *Biometrical Journal* **58**(3): 652–673.

Sauerbrei, W., Boulesteix, A.-L. and Binder, H. (2011). Stability investigations of multivariable regression models derived from low-and high-dimensional data, *Journal of Biopharmaceutical Statistics* **21**(6): 1206–1231.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model, *Statistics in Medicine* **11**(16): 2093–2109.

Shao, J. and Wu, C. J. (1989). A general theory for jackknife variance estimation, *The Annals of Statistics* **17**(3): 1176–1197.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. et al. (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell* **1**(2): 203–209.

Slawski, M., Daumer, M. and Boulesteix, A.-L. (2008). CMA–a comprehensive Bioconductor package for supervised classification with high dimensional data, *BMC Bioinformatics* **9**: 439.

Smyth, G. K. (2005). Limma: linear models for microarray data, *in* R. Gentleman, V. Carey, W. Huber, R. Irizarry and S. Dudoit (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, Springer, pp. 397–420.

Steck, H. and Jaakkola, T. S. (2003). Bias-corrected bootstrap and model uncertainty, *in* S. Thrun, L. Saul and B. Schölkopf (eds), *Advances in Neural Information Processing Systems 16*, Cambridge, MA: MIT Press.

Steidl, C., Lee, T., Shah, S. P., Farinha, P., Han, G., Nayar, T., Delaney, A., Jones, S. J., Iqbal, J., Weisenburger, D. D. et al. (2010). Tumor-associated macrophages and survival in classic Hodgkin's lymphoma, *New England Journal of Medicine* **362**(10): 875–885.

Strasser, H. (2000). Data compression and statistical inference, *Multivariate Statistics: Proceedings of the 6th Tartu Conference, Tartu, Estonia, 19-22 August 1999*, Walter de Gruyter, pp. 151–171.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A. (2008). Conditional variable importance for random forests, *BMC Bioinformatics* **9**: 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics* **8**: 25.

Strobl, C., Malley, J. and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests, *Psychological Methods* **14**(4): 323–348.

Strobl, C. and Zeileis, A. (2008). Danger: High power! – Exploring the statistical properties of a test for random forest variable importance, *Technical Report 17*, Department of Statistics, University of Munich.

Tan, A. C. and Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification, *Applied Bioinformatics* **2**(3 Suppl): S75–S83.

Tang, R., Sinnwell, J. P., Li, J., Rider, D. N., de Andrade, M. and Biernacka, J. M. (2009). Identification of genes and haplotypes that predict rheumatoid arthritis using random forests, *BMC Proceedings* **3**(Suppl 7): S68.

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M. and van Hijum, S. A. (2012). Data mining in the life sciences with random forest: a walk in the park or lost in the jungle?, *Briefings in Bioinformatics* **14**(3): 315–326.

Tutz, G. (2011). *Regression for categorical data*, Cambridge University Press, Cambridge.

van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature* **415**(6871): 530–536.

Waegeman, W., De Baets, B. and Boullart, L. (2008). ROC analysis in ordinal regression learning, *Pattern Recognition Letters* **29**(1): 1–9.

Wagenmakers, E.-J., Farrell, S. and Ratcliff, R. (2004). Naïve nonparametric bootstrap model weights are biased, *Biometrics* **60**(1): 281–283.

Wang-Sattler, R., Yu, Z., Herder, C., Messias, A. C., Floegel, A., He, Y., Heim, K., Campillos, M., Holzapfel, C., Thorand, B. et al. (2012). Novel biomarkers for pre-diabetes identified by metabolomics, *Molecular Systems Biology* **8**: 615.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis, *The Annals of Statistics* **14**(4): 1261–1295.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P. et al. (2012). Human gut microbiome viewed across age and geography, *Nature* **486**(7402): 222–227.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.2011, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 12.01.2016

Silke Janitza