LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Silke Janitza, Harald Binder, Anne-Laure Boulesteix

Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications

Technical Report Number 163, 2014
Department of Statistics
University of Munich

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

Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications

Silke Janitza[1]* Harald Binder[2] Anne-Laure Boulesteix[1]

June 27, 2014

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany.

[2] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center Johannes Gutenberg University Mainz, Obere Zahlbacher Str. 69, 55131 Mainz, Germany.

Abstract

The bootstrap method has become a widely used tool that has been applied in diverse areas where results based on asymptotic theory are scarce. It can be applied for example for assessing the variance of a statistic, a quantile of interest or for significance testing by resampling from the null hypothesis. Recently some approaches have been suggested in the biometrical field where hypothesis testing or model selection is performed on a bootstrap sample as if it was the original sample. From the literature, however, there is evidence that these procedures might lead to more significant results or overcomplex models, respectively, when ignoring that the bootstrap sample is not a direct realization of the true underlying distribution. We explain why this is the case and illustrate that tests on bootstrap samples do not provide valid p-values, using the Z-test and likelihood ratio test as examples. We also illustrate that information criteria when computed based on bootstrap samples are not reliable, as suggested by known theory. Furthermore, we revisit four approaches in light of these considerations: estimation of the p-value distribution, model complexity selection, variable inclusion frequencies, and model averaging. Using simulation studies and evidence from the literature we demonstrate that these approaches might give misleading conclusions and discuss possible solutions to this problem.

Keywords: Bootstrap; Bootstrap test statistic; Model selection; Model stability; Tests on bootstrap samples.

*Corresponding author. Email: janitza@ibe.med.uni-muenchen.de.

1 Introduction

The bootstrap method proposed by Efron (1979) has become a popular tool that is applied in diverse areas. It is becoming more and more widely used, e.g., as indicated by a larger number of textbooks (Chernick; 2011; Manly; 2006; Good; 2005; Davison; 1997). Bootstrapped statistics can for example be used to compute the variability of the statistic, a quantile of interest, a confidence interval or simply approximate the whole underlying distribution of the statistic. When considering biostatistical model building, there are however some results indicating problematic properties of the bootstrap, in particular when using approaches based on p-values.

The problem of deriving p-values from bootstrap samples is the fact that, even if the null hypothesis was true, when drawing a bootstrap sample from the original sample one is not sampling from a distribution where the null hypothesis holds. The empirical distribution is always slightly different from the true underlying distribution, and the estimated value of the parameter of interest computed from the observed sample randomly varies around the value of the population parameter but is never exactly equal to this true value.

This problem has already been reported in literature (Bollen and Stine; 1992; Strobl et al.; 2007) and has also been shown to be relevant when computing information criteria based on bootstrap samples (Steck and Jaakkola; 2003; Wagenmakers et al.; 2004). Suggestions have been made – for example Bollen and Stine (1992) derive a transformation of the empirical sample in the context of goodness-of-fit measures in structural equation models – to guarantee that one draws from a sample for which the null hypothesis holds. A straightforward and easy-to-use approach which in principle can be universally applied is to replace bootstrap sampling by subsampling, i.e. to draw from the original sample without replacement instead of drawing with replacement. This has for example been suggested by Strobl et al. (2007) in the context of the random forest method. To address the considered issue in the context of information criteria for graphical models, Steck and Jaakkola (2003) proposed a corrected version of the AIC.

However, the literature on problematic properties of the bootstrap when used for deriving p-values or for model building purposes is sparse and unfortunately, these problematic results are scattered over different scientific areas and research communities. Thus no comprehensive guidance on potential problems of using the bootstrap for model building is available for biostatistical applications. This will be addressed in the following. In this paper we collect theoretical and empirical evidence for these problems from studies published in heterogenous literature probably largely unknown to the biometrical community. Furthermore we extend these results from the literature both from a theoretical and empirical point of view (through own simulations) to provide better insight into the considered issues and their practical consequences in biometrical applications. All `R` codes implementing our simulation studies are provided on `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/janitza/index.html` for reproducibility purposes.

This paper is divided into two parts. In the first part of this paper we extend evidence from the literature for the increased type I error of tests performed on bootstrap samples, using as examples the Z-test and the likelihood ratio (LR) test. We furthermore motivate that this problem is also present when computing information criteria like the AIC from a bootstrap sample. Our considerations build on the previously reported results by Bollen and Stine (1992), Steck and Jaakkola (2003) and Wagenmakers et al. (2004). The previously published literature on this problem is very specific to a certain context and addresses readers from a very specific field.

In order to make this problem more accessible to applied (bio)statisticians we formulate it in a general and less technical manner and also provide empirical evidence. In the second part of this paper we discuss a selection of existing approaches used in biometrical applications, where a statistical test or model selection is – sometimes subconsciously – performed on a bootstrap sample. In the first approach the bootstrap is used for assessing the variability of p-values. The second approach deals with the selection of the optimal tuning parameter, where tuning parameter selection is done using an information criterion or cross-validation on a bootstrap sample in order to use the observations that were not drawn into the bootstrap sample for model evaluation. The third approach deals with model selection performed on each bootstrap sample that is commonly done in biometrical applications for assessing the stability of a model selection procedure. The last approach is a model averaging procedure in which AIC values are computed from models that were fit on bootstrap samples. Using simulation studies and evidence from the literature we illustrate that such bootstrap approaches might lead to misleading conclusions and provide the readers with tentative practical recommendations on how to proceed in these situations.

2 Theoretical Considerations

2.1 Principle of the bootstrap

Let θ be the parameter of interest from the true underlying distribution F. In practical applications there is usually one sample available that was drawn from F which can be used to compute an estimate $\hat{\theta}$ for the true population parameter θ. Since F is unknown it is not possible to draw additional samples from F in order to compute several estimates for the statistic θ. This is the point where the bootstrap comes into play. The idea is that in the "bootstrap world" one knows the true distribution F^* such that one can generate as many samples from F^* as desired. From each sample a bootstrapped statistic $\hat{\theta}^*$ can be computed which is an approximation of the estimate $\hat{\theta}$. F^* is chosen based on the available sample. In nonparametric bootstrap, which we are considering in this paper, it is simply the empirical distribution $\hat{F}$. In the case of parametric bootstrap one assumes a distribution type (e.g., a normal distribution) from which the data could come and estimates its parameters using the observed sample. For many statistics it can be shown that the behavior of the bootstrapped statistic $\hat{\theta}^*$ is a good approximation of the behaviour of the statistic $\hat{\theta}$ derived from the original sample. In this paper we are interested in the reliability of bootstrapped statistics if a test statistic or the corresponding p-value is the statistic of interest. In the following we will explore this in detail for the special case of the Z-test statistic. We derive theoretical results and also give empirical evidence. For the LR test this issue has already been addressed by Bollen and Stine (1992) and their results are briefly sketched here and underlined using empirical results.

2.2 Hypothesis testing on bootstrap samples

Z-test

For the moment let us assume that we want to perform a Z-test using data from the original distribution F, which is the normal distribution, and that the variance σ^2 is known. Let $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$ be independent and identically distributed (iid) random variables and $x_i, i = 1, \ldots, n$ the corresponding realizations. This set of realizations is termed "original sample" and denoted by $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$. The considered null hypothesis states that μ is equal to a predefined value μ_0. The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ has expectation μ and variance $\frac{\sigma^2}{n}$. The test statistic for the Z-test is then given by

$$Z = \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma}.$$

As known from theory the test statistic Z follows a standard normal distribution with $\mathrm{E}(Z) = \sqrt{n}\frac{\mu-\mu_0}{\sigma}$ and $\mathrm{Var}(Z) = 1$.

Now we derive the expectation and variance of Z when $\bar{x}$ is computed from a bootstrap sample $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)^\top$ that was drawn from the empirical distribution $\hat{F}$ of the original sample. The resulting test statistic is denoted by Z^* in the following. The expectation of Z^*, in which we are interested, will be denoted $\mathrm{E}(Z^*|F)$ to underline that this is the expected value we expect for the "real world" with true distribution F. The expectation of Z^* that is expected in the "bootstrap world" is termed $\mathrm{E}(Z^*|\hat{F})$. To derive $\mathrm{E}(Z^*|F)$ we make use of the law of iterated expectations which states that

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

We first compute the inner expected value $\mathrm{E}(Z^*|\hat{F})$. This is the value of Z^* that is expected in the "bootstrap world". It exactly equals the test statistic Z that is observed in the "real world" since in the "bootstrap world" we draw the samples from the empirical distribution $\hat{F}$ (Bollen and Stine; 1992). Taking the expectation of Z with respect to the "real world" gives $\mathrm{E}(Z|F) = \sqrt{n}\frac{\mu-\mu_0}{\sigma}$. Thus the expectations $E(Z|F)$ and $E(Z^*|F)$ are equal.

To derive the variance $\mathrm{Var}(Z^*|F)$ of Z^* we use the law of total variance which states that we can decompose the variance as follows:

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

Since $\mathrm{E}(Z^*|\hat{F}) = Z$ the first term in Eq. (1) reduces to $\mathrm{Var}(Z|F) = 1$. As far as the second term is concerned, the basic assumption underlying bootstrap estimation of the variance, which can be easily shown in the present simple special case (Davison; 1997), is that $\mathrm{Var}(Z^*|\hat{F})$ approximates $\mathrm{Var}(Z|F)$. Since $\mathrm{Var}(Z|F) = 1$ the second term in Eq. (1) becomes 1. Summing up both terms yields $\mathrm{Var}(Z^*|F) = 2$; the variance of Z^* is thus twice as large as the variance of Z. With this result it is proven that the test statistic Z^* that is computed from a bootstrap sample $\boldsymbol{x}^*$ does not follow the same distribution as the test statistic Z that is computed from the original sample $\boldsymbol{x}$.

This can also be observed from empirical results in a small simulation study. For computing Z and Z^* we draw $n = 1000$ independent observations from the standard normal distribution. We then draw a bootstrap sample out of this original sample and compute the test statistic for a Z-test with null hypothesis $H_0 : \mu = 0$ from both original and bootstrap samples. This procedure is repeated 500000 times, yielding 500000 values of Z and Z^*, respectively. Figure 1 shows the resulting empirical density function of Z and Z^*. As expected from theory the distribution of the test statistic Z coincides with the standard normal distribution since the respective lines in Figure 1 show a perfect coverage. 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 value seems to be equal. The empirical expectation of Z and Z^* are both close to the value zero with values -0.0010 and 0.0018, respectively. In contrast to that the empirical variance of Z^* is, at 2.0011, larger by factor 2 than the variance of Z, which is, at 1.0009, very close to the variance of the standard normal distribution. These empirical results are thus in line with our theoretical derivations.

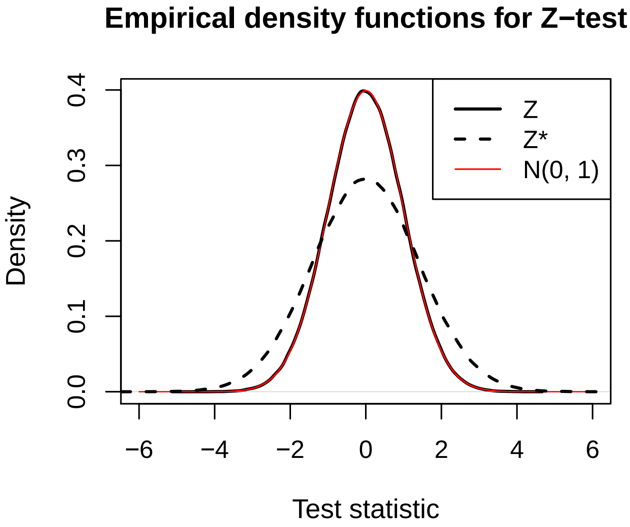

Figure 1: Empirical density functions for test statistics Z (solid black line) and Z^* (dashed black line) of the Z-test. The density of the standard normal distribution is indicated by the red line.

One can now ask what happens if the test statistic Z^* computed from bootstrap samples is used in combination with a significance threshold taken from the null distribution of the original test statistic Z to compute a p-value. To answer this question we consider the distribution of Z and Z^* if the null hypothesis $H_0 : \mu = \mu_0$ holds in the "real world". According to our results presented above, under the null hypothesis Z has expectation 0 and variance 1 (thus following a standard normal distribution) while Z^* has expectation 0 and variance 2. Using this information one can derive the actual type I error for the test with test statistic Z^* and significance threshold taken from the test distribution for Z. Since Z follows a standard normal distribution under the null hypothesis, the significance threshold is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. It can be easily derived that when using the significance threshold $z_{1-\frac{\alpha}{2}}$ for a two-sided test with test statistic Z^*, the actual 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 significance threshold z_α ($z_{1-\alpha}$) one has an actual type I error of $\Phi(\frac{1}{\sqrt{2}} z_\alpha)$ (and $1 - \Phi(\frac{1}{\sqrt{2}} z_{1-\alpha})$, respectively). Table 1 shows examples for the supposed and actual type I error when performing Z-tests using test statistic Z^* and significance thresholds for Z. It can be seen that the actual type I error is substantially increased when performing Z-tests on bootstrap 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 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

Supposed type I error	Actual type I error	
	two-sided Z-test	one-sided Z-test
0.10	0.24	0.18
0.05	0.17	0.12
0.01	0.07	0.05

Table 1: Supposed and actual type I error when performing Z-tests using test statistic Z^* computed from bootstrap sample with significance thresholds taken from the standard normal distribution.

difference in log-likelihoods:

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

The test statistic T asymptotically follows a non-central χ^2-distribution with df degrees of freedom that are obtained from 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 $\mathrm{AE}(T) = df + \kappa$ and the asymptotic variance is $\mathrm{AVar}(T) = 2df + 4\kappa$. 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 $\mathrm{AE}(T) = df$ and asymptotic variance $\mathrm{AVar}(T) = 2df$. This is proven to hold for models derived from the original sample $\boldsymbol{x}$.

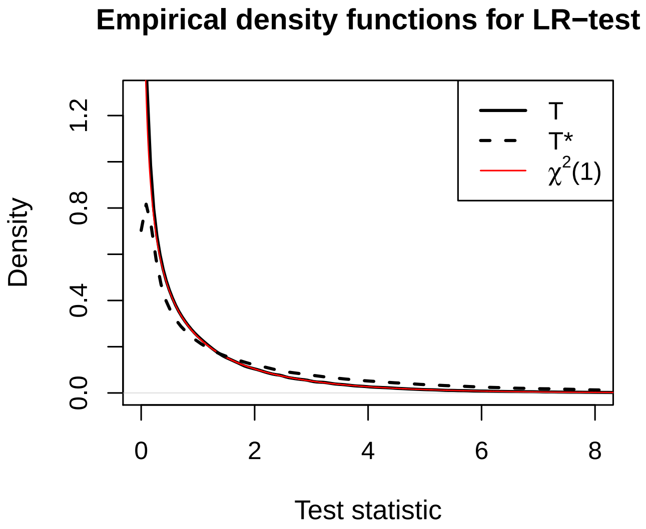

Figure 2: Empirical density functions for test statistics T (solid black line) and T^* (dashed black line) of the LR test with 1 degree of freedom. The density of the $\chi^2(1)$-distribution is indicated by the red line.

Bollen and Stine (1992) gave an approximation for the asymptotic expectation of the test statistic T^* that is derived from bootstrap sample $\boldsymbol{x}^*$. They report it being twice as large as the asymptotic expectation of T in the original sample. They also report the asymptotic variance of T^* to be larger than the asymptotic variance of T. These theoretical results are in line with our empirical results from a simulation study where we used a very large sample size of $n = 100000$. We draw predictor values $X_i \sim N(0, 1)$ and independently of the predictor values we draw response values $Y_i \sim N(0, 1)$ for observations $i = 1, \ldots, n$. Subsequently a bootstrap sample is drawn from this original sample. A LR test with one degree of freedom is performed on the original sample and on the bootstrap sample to test if the linear regression model including predictor X gives a better model fit than the intercept model. From this test we obtain test statistics T for the original sample and T^* for the bootstrap sample. The data generation and computation of the test statistics are repeated 500000 times in order to obtain empirical distributions of T and T^*.

Figure 2 shows the empirical density functions of T and T^*. The distribution of T approximates the χ^2-distribution with 1 degree of freedom very well (the respective lines in Figure 2 coincide) which indicates that the number of observations was chosen high enough. It is remarkable that the distribution of T^* noticeably deviates from the χ^2-distribution. It has a much higher variability so that the probability mass in the tail is larger compared to that of T. To quantify the discrepancy between the empirical distributions of T and T^* we compute the empirical expectation and variance of T and T^*. While the empirical expected value of T is, at 0.9977, very close to the true asymptotic expectation of 1, the empirical value of T^* is 2.0005 and is thus approximately twice as large, as suggested by the approximation provided by Bollen and Stine (1992). The empirical variance of T is, at 1.9819, also close to the theoretical approximate variance of 2. The variance of T^* in contrast is with a value of 8.0140 higher by a factor of 4. From these results it is obvious that the type I error is increased when performing a LR test on bootstrap samples using the critical values from a χ^2-distribution.

2.3 Information criteria based on bootstrap samples

Information criteria like the AIC and BIC are similarly affected when derived from models built on bootstrap samples. In the context of graphical models Steck and Jaakkola (2003) proved that the bootstrapped information criteria systematically deviate from information criteria derived from original samples. We illustrate their findings using the correspondence between the likelihood ratio test statistic (LRT) and information criteria for nested models. For this purpose we assume that we aim to compare two nested models using the AIC and that the models differ in the inclusion of only one parameter (similar considerations can be made in the case of nested models differing by the inclusion of more than one parameter). The definition of the AIC is:

$$\mathrm{AIC} = -2\log(L) + 2p, \tag{3}$$

where L denotes the likelihood and p denotes the number of parameters included in the model. 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 LRT on one degree of freedom can be expressed in terms of AIC_0 and AIC_1 (cf. Chapter 6.9.3 in Burnham and Anderson; 2002):

$$\mathrm{LRT} = \mathrm{AIC}_0 - \mathrm{AIC}_1 + 2. \tag{4}$$

From Eq. (4) we see that if both models fit the data equally well according to the AIC (i.e. $\mathrm{AIC}_0 = \mathrm{AIC}_1$), we have $\mathrm{LRT} = 2$. Further, the unrestricted model is chosen over the submodel if its AIC is smaller, corresponding to $\mathrm{AIC}_0 - \mathrm{AIC}_1 > 0$ and, according to Eq. (4), $\mathrm{LRT} > 2$. In contrast, the submodel is chosen if it has a smaller AIC value: $\mathrm{AIC}_0 - \mathrm{AIC}_1 < 0$ corresponding to $\mathrm{LRT} < 2$. These considerations illustrate that in the case of two nested models one can also use the value of the LRT to decide which of the models is better in terms of the AIC; values for the LRT below 2 are in favor of the submodel and values above 2 indicate that the unrestricted model is better. Both models are considered equally good in terms of the AIC if the LRT takes the value 2. As shown in the first part of this section, bootstrapped LRT values are not valid. Due to the correspondence between the LRT and the AIC under the specific setting of nested models it is proven that bootstrapped information criteria like the AIC are thus not valid as well.

Steck and Jaakkola (2003) show that information criteria like the BIC and the AIC used for structure learning for Bayesian networks are not valid when computed from bootstrap samples. They give a thorough technical description which we will not present here. In particular, they show that the log-likelihood terms in bootstrapped information criteria are not valid and should be decreased by $\frac{1}{2}p$. By subtracting this term from the log-likelihood – or equivalently by adding p to the information criterion – one obtains an almost unbiased information criterion. According to Steck and Jaakkola (2003) in order to learn the structure of graphical models using the bootstrap one can use a bias corrected version of the AIC, denoted by AIC^{BC}. This is defined as:

$$\begin{aligned}\text{AIC}^{BC} &= -2\left(\log(L) - \frac{1}{2}p\right) + 2p \\ &= -2\log(L) + 3p. \end{aligned} \tag{5}$$

As noted by the authors, since the bias originates from the log-likelihood the same bias correction term can be applied to any information criterion that is computed from the log-likelihood and a penalty term.

The studies by Steck and Jaakkola (2003) show that more complex models, in terms of included parameters, have actually too high a likelihood. This fact is ignored when computing the traditional AIC for models built from bootstrap samples: When selecting the model that was built from a bootstrap sample by using the uncorrected AIC, one would select a more complex model on average. This model would possibly not have been selected if the model has been built from the original sample. The bias corrected version of the AIC in Eq. (5) for learning graphical models, if compared to the original version of the AIC given in Eq. (3), suggests that in a bootstrap sample the number of parameters should be more strongly penalized (by factor 3) than in the original sample (factor 2) in order to obtain comparable results for the AIC value computed from bootstrap and original samples in the context of graphical models.

3 Practical Consequences

In this section we discuss four approaches based on the bootstrap which use the information of p-values or information criteria that were derived from bootstrap samples. Keeping in mind the results presented in the previous section, we illustrate that these approaches might lead to misinterpretations and provide evidence through simulation studies.

3.1 Variability of p-values

In their paper Boos and Stefanski (2011) introduce a bootstrap approach to explore the variability of p-values. In this approach a p-value is computed from each bootstrap sample such that the variability of p-values (or -$\log_{10}$(p-value)) can be assessed. However, this approach is not valid; the problem in computing the variability of p-values from bootstrap samples is that the p-value distribution is different when computed on bootstrap samples. This has been theoretically motivated in important special cases (Z-test and LR test) in the first part of this paper and will be confirmed by the use of simulation studies here. For a Z-test we independently draw $n = 1000$ observations from the standard normal distribution for testing the null hypothesis that the population mean equals zero. Under the alternative we draw from $N(0.08, 1)$. A bootstrap sample is generated by drawing from the original data and a Z-test is performed separately for the original sample and the bootstrap sample. This process is repeated 5000 times.

For the LR test p metric predictor variables $x_{i1}, \ldots, x_{ip}$ are independently drawn for $i = 1, \ldots, 1000$ from a multivariate normal distribution with expected value $\boldsymbol{\mu} = (0, \ldots, 0)^\top \in \mathbb{R}^p$ and variance $\boldsymbol{I}_p$ corresponding to the identity matrix of dimension p. The response variable Y_i is generated according to the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$. The global null hypothesis states that none of the p predictors is associated with the response, i.e. $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ and the alternative hypothesis is that at least one of the coefficients is associated, i.e. $H_1 : \beta_j \neq 0$ for at least one $j \in \{1, \ldots, p\}$. The corresponding LR test compares the likelihood of the submodel L_0 containing only the intercept to the likelihood L_1 of the model containing all predictor variables. If the null hypothesis is true the LR test statistic (2) follows a central χ^2-square distribution with p degrees of freedom. In our simulations all beta coefficients are set to the value zero for the setting under the null hypothesis and to the value 0.02 if the alternative hypothesis is true. For simulations on the LR test we perform several simulation settings with different numbers of predictor variables p. These show slight to severe differences to the true p-value variability when computed on bootstrap samples. Here we show only the results for the setting with $p = 10$ predictor variables (corresponding to a LR test with 10 degrees of freedom) in which the discrepancy with the true p-value variability is rather extreme.

Figure 3 shows the distribution of p-values for the Z-test and the LR test: it is obvious that the p-value distribution derived from bootstrap samples is not a good approximation of the p-value distribution that is obtained for original samples. The corresponding standard deviations of the p-value (or -$\log_{10}$(p-value)) computed on bootstrap samples do not reflect the true p-value variability in our studies, neither under the null hypothesis nor under the alternative hypothesis. In practical applications one should thus be aware that misleading results might be obtained when deriving p-values from bootstrap samples or other statistics that are based on p-values, such as the p-value variability. Subsampling is not a reasonable alternative here if more than estimation of the p-value distribution under the null hypothesis and type I error control is wanted. Figure 4 shows the corresponding results for the Z-test and the LR test when the test is performed on subsamples of size $0.632n$. The value 0.632 was chosen because this is the expected number of unique observations in a bootstrap sample. One can see that tests performed on subsamples do preserve the α-level but one obtains higher p-values under the alternative hypothesis, which is attributable to the decreased statistical power. Tests performed on subsamples thus do not reflect the p-value variability in original samples, either.

3.2 Tuning parameter selection

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

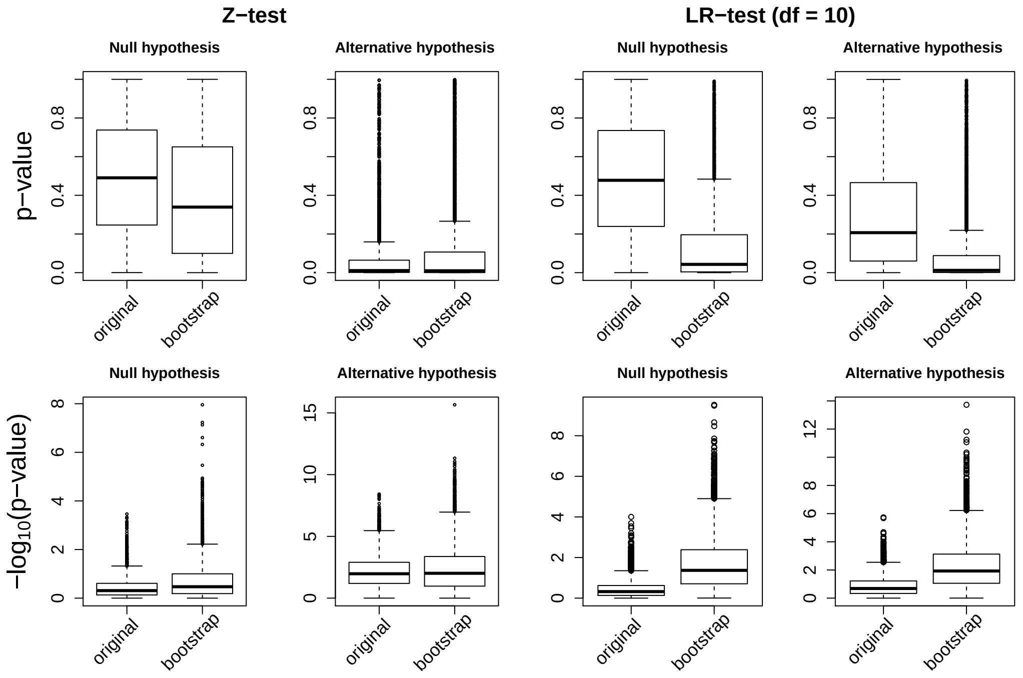

Figure 3: p-value distribution for Z-test (left two columns) and LR test with 10 degrees of freedom (right two columns) computed on 5000 original samples (left boxplot) and 5000 bootstrap samples (right boxplot) under the null hypothesis and under the alternative hypothesis.

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 (often termed "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 et al.; 2007), the optimal value for the tuning parameter is often determined by using information criteria or through application of an internal cross-validation procedure. In the following we illustrate that when performing either of these procedures on a bootstrap sample a tuning parameter is selected which leads to (at least slightly) more complex models. We will also investigate the use of the subsampling procedure for splitting the data into a training and a test set to see whether the subsampling procedure could be used to solve this problem. We first consider the case where the tuning parameter is selected by the use of an information criterion and then investigate the use of cross-validation procedures for tuning parameter selection. As an example we here consider the selection of the optimal number of boosting steps for gradient boosting algorithms.

Selection via information criteria

As illustrated in the first part of this paper, when computing the AIC for models that were fit on bootstrap samples and deciding for the model with the minimal AIC, one would select a more complex model, i.e. more parameters. In the specific context of gradient boosting algorithms more complex models are obtained when performing a larger number of boosting steps. Accordingly,

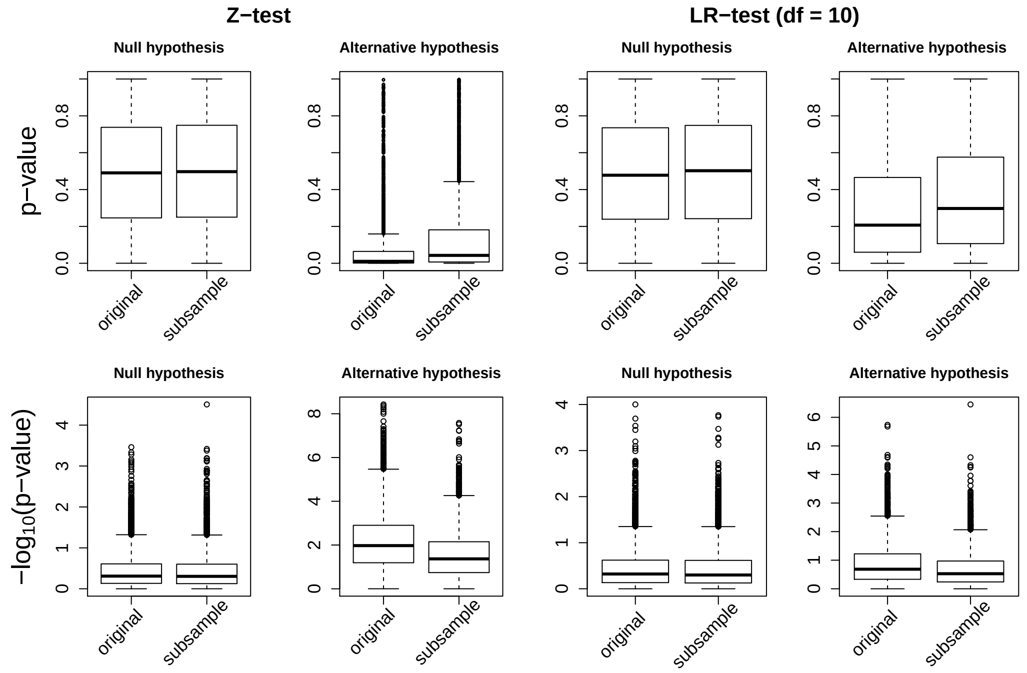

Figure 4: p-value distribution for Z-test (left two columns) and LR test with 10 degrees of freedom (right two columns) computed on 5000 original samples (left boxplot) and 5000 subsamples (right boxplot) under the null hypothesis and under the alternative hypothesis.

when determining the number of boosting steps that leads to the minimal AIC value, one would select a higher number of boosting steps. This can also be seen in simulation studies where the AIC is used to determine the optimal number of boosting steps. In our simulation studies we compare the optimal number of boosting steps for which the AIC is minimized on the original data (i.e., we use the whole data as training set), on a bootstrap sample and on a subsample of size $0.632n$. The data generating process is the same as that described by Binder and Schumacher (2008) for the simulation study on binary response gradient boosting. Data is simulated for the uncorrelated setting, where $p \in \{200, 1000, 5000\}$ predictors are 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 medium effects), as per the simulation studies by Binder and Schumacher (2008). The binary response value for an observation with covariates $\boldsymbol{x}_i$ is simulated from a binomial distribution with success probability $\pi_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$. We determine the optimal number of boosting steps on the original data, a bootstrap sample and a subsample. This is repeated 1000 times.

Figure 5 shows the optimal number of boosting steps for the setting with weak effects. The results for the setting with moderate effects are comparable and are thus not shown. The results of the simulation studies support our hypothesis that a higher number of boosting steps, or equivalently, a higher complexity of gradient boosting models, is chosen when performing tuning

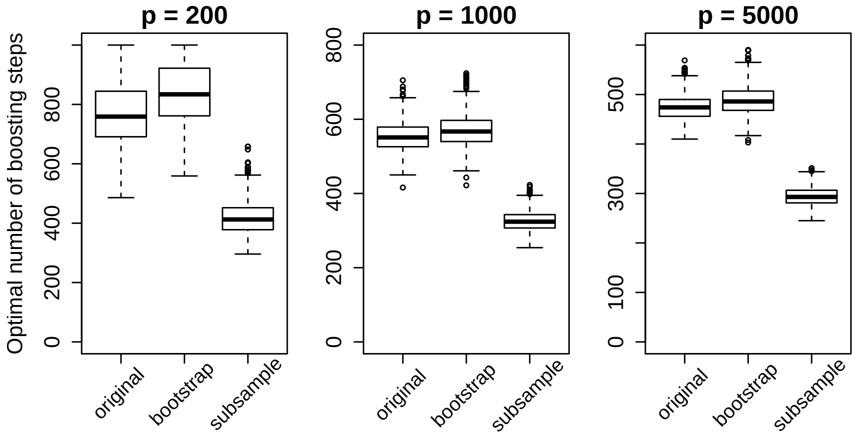

Figure 5: Optimal number of boosting steps selected via AIC for binary response gradient boosting in 1000 original samples, bootstrap samples and subsamples for the setting with weak effects ($c_e = 1$).

parameter selection on bootstrap samples. Interestingly, however, the amount of overcomplexity induced by the bootstrap is negligible in our studies since only a marginally higher number of boosting steps is chosen. The discrepancy in selected boosting steps for original samples and for subsamples is much more pronounced than the difference in selected boosting steps for original samples and bootstrap samples. In contrast to the bootstrap, when using subsamples a substantially smaller number of boosting steps is selected. To conclude, the results have shown that tuning parameter selection via AIC on bootstrap samples results in slightly more complex boosting models. Though, the amount of overcomplexity seems to be negligible in our studies. Using subsampling for splitting the data into training and test sets, in contrast, yields considerably smaller numbers of boosting steps that lead to considerably sparser models. Thus, although the bootstrap systematically induces a higher complexity, in our studies the induced overcomplexity is negligible and comes much closer to the true model complexity than the subsampling approach that chooses considerably more simplistic models.

Selection via cross-validation

Alternatively, instead of using information criteria one may use cross-validation for selecting the optimal value for a tuning parameter. 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 higher 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. The preference for overcomplex models can be explained as follows: In a bootstrap sample the same original observation can occur several times. Thus when performing cross-validation on a bootstrap sample the same original observation may be present in the training as well as in the test set. If a model is evaluated on observations that were already used for fitting the model, more complex models might imply a better fit. However, these have poor predictive accuracy on new data. A solution to this problem 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. Here we consider a different approach for preventing an overlap, in which all duplications of the same observation in a bootstrap sample are regarded as one unit and randomly split the units – instead of the observations – into k sets, in which each set contains an equal number of units. To investigate if this latter approach, or subsampling, might be used for tuning parameter selection via k-fold cross-validation on bootstrap samples, we perform a simulation study. The data is generated the same way as described in the preceding paragraph. We compute the optimal number of boosting steps for which the cross-validated empirical loss is minimized when 5-fold cross-validation is performed on original samples, on bootstrap samples allowing training and test sets to overlap, on bootstrap samples not allowing training and test sets to overlap and on subsamples of size $0.632n$ containing no duplicated observations. The results for 1000 original samples, bootstrap samples and subsamples are shown in Figure 6 for the setting with weak effects ($c_e = 1$).

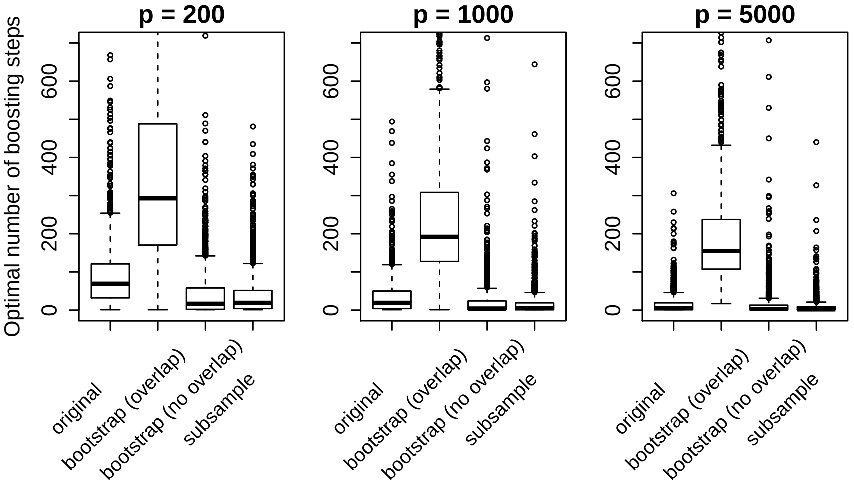

Figure 6: Optimal number of boosting steps selected by 5-fold cross-validation for binary response gradient boosting in 1000 original samples, bootstrap samples (with and without allowing training and test sets from 5-fold cross-validation to overlap), and subsamples, for the setting with weak effects ($c_e = 1$).

As already seen in the studies by Binder and Schumacher (2008), a considerably higher number of boosting steps is obtained when performing cross-validation on classical bootstrap samples (where training and test sets may overlap) compared to original samples. Interestingly, when performing cross-validation on bootstrap samples with the restriction that training and test sets cannot overlap there is no tendency towards a higher number of boosting steps. More precisely, the median number of boosting steps is much lower than for original samples and comes very 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 samples (i.e., 63.2% of the original sample), one might hypothesize that the information content in a bootstrap sample and in the subsample is approximately the same and that duplicated observations do not affect the boosting algorithm as long as training and test sets do not overlap. The results suggest that the larger number of boosting steps for the classical bootstrap approach, or equivalently, the higher complexity of gradient boosting models, results from the overlap of training and tests sets when performing cross-validation on bootstrap samples.

We conclude from these studies that with the classical bootstrap approach too complex models are promoted, as already seen in the studies by Binder and Schumacher (2008). Performing cross-validation on ordinary bootstrap samples for choosing the optimal number of boosting steps is thus not recommendable. With the modified bootstrap approach (that prevents an overlap of training and test sets) and the subsampling approach sparser models are chosen compared to models derived on original samples but the difference is rather small in our studies indicating that both approaches may be used for tuning parameter selection.

3.3 Model stability issues

A major field where several bootstrap-based approaches have been developed in the past is model selection for multivariable regression. It is well known that model selection strategies suffer from instability. The bootstrap method is an attractive tool for stability investigations since it offers the possibility to obtain slightly perturbed datasets from the original data. Chen and George (1985), Altman and Andersen (1989) and Sauerbrei and Schumacher (1992) for example proposed performing model selection on bootstrap samples. With each bootstrap sample one obtains a model, in this way enabling stability investigations. The LR test can for example be used to decide on the inclusion or exclusion of a variable in each step of the model selection procedure. Other criteria like the AIC or BIC could also be used. Using this approach one can investigate the importance of variables via so-called *bootstrap inclusion frequencies*. These correspond to the relative frequency a variable is included in the resulting models. Roughly speaking a variable with a high bootstrap inclusion frequency indicates a variable with a high prognostic importance, while a variable with a low bootstrap inclusion frequency is thought to have only little prognostic importance (see Sauerbrei and Schumacher; 1992, for more detailed information, e.g., on the handling of correlated variables). The bootstrap-based approach proposed by Sauerbrei and Schumacher (1992) can be seen as a reasonable tool to assess the importance of predictor variables while addressing the instability of classical approaches.

However, recent studies showed that model selection performed on bootstrap samples results in models including more predictor variables and that categorical predictors with many categories are more often included than predictors with fewer categories or metric predictors when compared to samples drawn from the original distribution (Rospleszcz et al.; 2014). The inclusion of more predictor variables on bootstrap samples results from the fact that the LR test has an increased type I error when performed on bootstrap samples, thus finding more significant associations during the model selection process and including more predictor variables in the model. This issue is thus directly related to our considerations from Section 2. If the inflation were independent of the type of predictor (e.g., categorical or metric) this overcomplexity would not be a problem, since the significance level can be seen as more or less arbitrary in the context of model selection for multivariable regression and one could just decrease the significance level to achieve a less complex model. However, the type I error increase depends on the degrees of freedom of the LR test, which are determined by the number of parameters being tested. The more categories a variable has, the more degrees of freedom the test has and the more increased is the type I error. In the context of the χ^2-test, this result was already noted by Strobl et al. (2007), who provided evidence via simulation studies. In their studies the p-value distribution computed from χ^2-tests conducted on bootstrap samples is more skewed towards smaller values the more categories a predictor has. When using the described model selection approach on bootstrap samples for data with different types of predictors, it might happen that categorical predictors with many categories are preferentially selected due to a more extreme increase in type I error. Bootstrap inclusion frequencies might not be reliable in reflecting the relative importance of predictors if different types of predictors are present in the data since categorical predictors with many categories have systematically higher bootstrap inclusion frequencies than categorical predictors with fewer categories and metric predictors even under the null hypothesis, as documented by Rospleszcz et al. (2014). A solution to this problem is to perform the described approach on subsamples instead of bootstrap samples. The resulting "subsample inclusion frequencies" can be computed likewise and were shown to reliably reflect the importance of predictors (Rospleszcz et al.; 2014).

3.4 Model averaging

A model averaging method based on the bootstrap was suggested by Buckland et al. (1997) (see also Burnham and Anderson; 2002). In model averaging inference regarding a quantity of interest is not only based on one best model, determined for example by the minimal AIC, but on a set of plausible models, this way incorporating model uncertainty into inference. A quantity of interest might for example be a regression parameter or a predicted value. In model averaging, the quantity of interest is averaged over several plausible models where model k has a weight w_k that reflects its plausibility. The estimated quantity of interest averaged over all K models is then given by:

$$\hat{\theta} = \sum_{k=1}^{K} w_k \hat{\theta}_k,$$

where $\hat{\theta}_k$ is the model-specific quantity of interest for model k.

Buckland et al. (1997) and Burnham and Anderson (2002) introduce two approaches in which the bootstrap is used to determine weights w_k for models $k = 1, \dots, K$. One approach, here termed the *selection frequency method*, consists of fitting all K candidate models to each bootstrap sample and selecting the most plausible model (determined for example by the smallest AIC) in each bootstrap sample. The weight for model k is then determined by the fraction of bootstrap samples in which model k had been selected. Another approach, called the *average weight method*, was proposed by Burnham and Anderson (2002). In this approach they use the so-called *Akaike weights* that are derived from the AIC values. A model k is attributed a weight

$$w_k = \frac{\exp(-\Delta_k/2)}{\sum_{i=1}^{K} \exp(-\Delta_i/2)}, \tag{6}$$

where $\Delta_k = \text{AIC}_k - \min(\text{AIC}_1, \dots, \text{AIC}_K)$ measures the distance between model k and the best model with the minimal AIC, where AIC_k denotes the AIC for model k. From Eq. (6) it can be seen that the bigger the distance Δ_k, the smaller is the weight w_k and thus the plausibility of model k (Burnham and Anderson; 2002, p. 75). These weights are computed separately on each bootstrap sample and finally averaged over all B bootstrap samples to obtain a weight for each model (Burnham and Anderson; 2002, p. 172).

By using simulation studies Wagenmakers et al. (2004) showed that both types of weights are biased when computed on bootstrap samples. Their argumentation is based on the correspondence between the LR test and the information criteria for nested models, as illustrated in Section 2.3. They showed that both weighting methods provide larger weights for more complex models when derived from bootstrap samples and concluded that, when using these model averaging approaches, the plausibility of more complex models would be overestimated and inference would be affected since parameter estimates are more variable for more complex models.

4 Discussion

Bootstrap procedures are widely used in biometry to solve problems that are difficult to address using asymptotic theory. They can be applied for example to assess the variance of a statistic, a quantile of interest or for significance testing by resampling from the null hypothesis. With the introduction of the bootstrap in 1979 more and more approaches based on the bootstrap have been developed. However, when performing hypothesis tests on bootstrap samples as if they were original samples the type I error is increased. Similarly, information criteria like the AIC or BIC computed from a bootstrap sample depart from that of original samples. Unfortunately, such problematic results are scattered over many papers and communities, and so far have not been comprehensively considered for biostatistical modeling. Evidence or indications for the increased type I error or the discrepancy in information criteria when computed on bootstrap samples can be found in Bollen and Stine (1992) in the context of structural equation models, in Strobl et al. (2007) in the context of random forest methodology, in Wagenmakers et al. (2004) in the context of model averaging and in Steck and Jaakkola (2003) in the context of graphical models. In addition in almost all cases the title of the work does not reveal that this kind of "bootstrap bias" is dealt with, which aggravates the problem.

In the first part of this article we have illustrated in a general context that there is a discrepancy in variance, and for some test statistics also in the expected value, when computing test statistics from original and from bootstrap samples. This discrepancy in variability (as well as the discrepancy in the expected value for some tests) leads to an increased type I error when performing hypothesis tests on a bootstrap sample. The discrepancy arises from the fact that when drawing bootstrap samples from the empirical distribution – presuming that no modification is made to the data beforehand, which is usually the case – one is not drawing from the true distribution, since the empirical distribution is never exactly equal to the true distribution, due to sampling variability. Through the relationship between the LR test and the AIC we have illustrated that information criteria are not reliable when computed from bootstrap samples, as well. This had been shown before by Steck and Jaakkola (2003) who also derived a bias-corrected version of information criteria for learning graphical models and this issue had also been reported by Wagenmakers et al. (2004) in a letter to the editor in the context of model averaging procedures.

For the vast majority of approaches that make use of bootstrapped p-values or information criteria, the practical consequences are unknown and remain to be investigated. In the second part of this article we outlined four bootstrap approaches from the biometrical field and discussed the consequences of a possible bias induced by the bootstrap in these applications. For the first two approaches we discussed possible consequences and provided evidence through simulation studies. For the latter two approaches we used evidence from the literature to illustrate that the bootstrap approaches might give misleading results. We also investigated possible alternative strategies to circumvent a bias induced by the bootstrap, such as subsampling, which has often been mentioned as a promising alternative to the bootstrap. A recent approach to stability selection which is based on subsampling has been introduced by Meinshausen and Bühlmann (2010). Their studies impressively show that subsampling is a powerful tool to investigate the stability of models in different contexts such as penalized likelihood estimation and graphical modeling. However, we have demonstrated that subsampling should not be regarded as an universally applicable alternative to the bootstrap. For selecting the optimal number of boosting steps via information criteria for example, with the subsampling procedure a considerably too small number of boosting steps was selected in our simulation studies and thus cannot be recommended if the aim is to investigate the distribution of model complexity parameters. This might potentially affect other uses, for example when prediction performance is to be investigated. For investigating the variability of p-values subsampling is not appropriate either, if more than type I error control is wanted. This makes clear that for some approaches subsampling might be an alternative to the bootstrap that gives reliable results while for other approaches it cannot be applied for obtaining reliable results.

Applied researchers should be careful when using approaches where hypothesis tests or information criteria are computed based on a bootstrap sample. If no investigations exist that indicate the reliability of a bootstrap approach, simulation studies are a helpful tool to investigate this. It is important to keep in mind that the bootstrap cannot be applied to any procedure as if it were the original sample. It might be advisable for methodologists to check the validity of their proposed bootstrap approaches by using simulation studies and comparing the results of their bootstrap approach to those that are obtained when using original samples from the true underlying distribution instead of bootstrap samples. In this way unexpected results can easily be discovered and adjustments may be made.

Supplementary material

`R` code implementing our simulation studies is available at `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/janitza/index.html`

References

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.

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.

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

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: an integral part of inference, *Biometrics* **53**(2): 603–618.

Bühlmann, P., Hothorn, T. et al. (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.

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. (2011). *Bootstrap methods: A guide for practitioners and researchers*, 2 edn, New York, John Wiley & Sons.

Davison, A. C. (1997). *Bootstrap methods and their application*, Vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *The Annals of Statistics* **7**: 1–26.

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

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses*, 3 edn, New York, Springer.

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.

Manly, B. F. (2006). *Randomization, bootstrap and Monte Carlo methods in biology*, 3 edn, Florida, CRC Press.

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

Rospleszcz, S., Janitza, S. and Boulesteix, A.-L. (2014). Categorical variables with many categories are preferentially selected in model selection procedures for multivariable regression models on bootstrap samples, *Technical Report 164*, Department of Statistics, University of Munich.

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.

Steck, H. and Jaakkola, T. S. (2003). Bias-corrected bootstrap and model uncertainty, *Advances in Neural Information Processing Systems*, number 16.

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**(1): 25.

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