ELSEVIER

Chemometrics and Intelligent Laboratory Systems 54 (2000) 35–52

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

The bootstrap: a tutorial

Ron Wehrens [a,*], Hein Putter [b], Lutgarde M.C. Buydens [a]

[a] *Laboratory of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, Netherlands*

[b] *Department of Medical Statistics, University Medical Centre, P.O. Box 9604, 2300 RC Leiden, Netherlands*

Received 8 August 2000; accepted 20 August 2000

Abstract

Bootstrap methods have gained wide acceptance and huge popularity in the field of applied statistics. The bootstrap is able to provide accurate answers in cases where other methods are simply not available, or where the usual approximations are invalid. The number of applications in chemistry, however, has been rather limited. One possible cause for this is the overwhelming number of techniques available. This tutorial aims to introduce the basic concepts of bootstrap methods, provide some guidance as to what bootstrap methods are appropriate in different situations, and illustrate several potential application areas in chemometrics by worked examples. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Bootstrap; Chemometrics; Confidence intervals; Error estimate

Contents

[*] Corresponding author. Tel.: +31-24-365-2053; fax: +31-24-365-2653.
E-mail address: rwehrens@sci.kun.nl (R. Wehrens).

PII: S0169-7439(00)00102-7

1. Introduction

The bootstrap [1–4] is a widely applicable computer-intensive statistical tool that may yield estimates that in other ways would be difficult to obtain. Despite its simple basic ideas and its broad applicability, the number of applications in chemistry is still limited. This tutorial aims to introduce the central ideas behind the bootstrap, to identify areas where it may be more successful than other methods, and to discuss some of its weaknesses.

In this tutorial, extensive use will be made of a data set obtained from the field of structure–activity relationships. In an investigation of co-crystallisation behaviour of small organic molecules with cephradine, an antibiotic, a data set of 120 compounds was tested. Sixty-eight of these compounds form crystalline complexes with cephradine, which allows for an easy isolation of the cephradine from the reaction mixture after synthesis. For obvious reasons, the co-crystallising compound should satisfy several additional criteria concerning efficiency, non-toxicity and costs. Ideally, one would like to be able to predict for a set of untested compounds which of them will form complexes [5,6].

Since the shape of the organic compounds is important (they should fit in the cavities in the crystal, created by the cephradine host molecules), two important descriptors are the molecular volume (calculated using the Van der Waals radii of the atoms) and the ellipsoidal volume. A large difference between these two indicates a non-spherical molecular shape. In Fig. 1, these two descriptors are plotted for the data set of 120 compounds. Several questions may be asked. Is there a relation between the molecular volume and the ellipsoidal volume for the complexating compounds on the one hand, and for the non-complexating compounds on the other? If so, are the two relations different? Or do the complexating compounds form a much more well-defined group than the non-complexating compounds? One should not forget that the non-complexating compounds are not a random selection, but were picked on the basis of chemical intuition as potentially complexating compounds.

If we concentrate first on the relation between molecular volume and ellipsoidal volume for the 68 complexating compounds, we can fit a regression line. Since both descriptors are of approximately the same precision, an orthogonal regression line, given

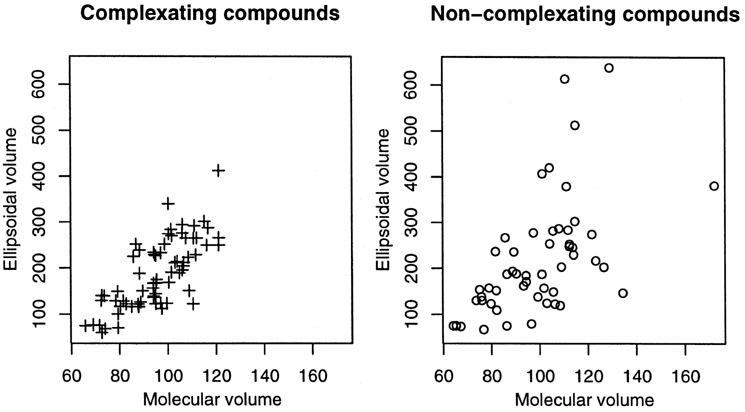

Fig. 1. Molecular volume plotted against ellipsoidal volume for the 120 compounds, where the 68 complexating compounds are plotted in the left panel, and the 52 non-complexating compounds in the right panel.

Table 1
The creation of bootstrap samples by sampling from the objects with replacement. Each bootstrap sample consists of 68 objects, some of which appear more than once

Object	Molecular volume	Ellipsoidal volume	Bootstrap sample 1	Bootstrap sample 2	...
1	79.3	100.4	object 3: (72.6, 60.2)	object 49: (115.1, 301.9)	...
2	79.4	70.4	object 50: (95.0, 226.8)	object 11: (97.1, 233.1)	...
3	72.6	60.2	object 20: (106.6, 211.9)	object 11: (97.1, 233.1)	...
⋮	⋮	⋮	⋮	⋮	
68	121.1	266.1	object 51: (95.1, 168.3)	object 63: (94.2, 139.4)	...
	$y = -497.3 + 7.18x$		$y = -453.8 + 6.87x$	$y = -388.7 + 5.98x$	...

by the first principal component, could be used [7]. The line obtained in this way is

$$y = -497.3 + 7.18\,x,$$

where x indicates the molecular volume, and y the ellipsoidal volume. Using classical techniques, it is not so easy to obtain standard errors or confidence intervals for the line parameters. Using the bootstrap, however, they are readily found. The main trick is that many new data sets (called bootstrap samples) are created from the original data set by sampling with replacement. In the regression example, bootstrap samples of 68 complexating compounds are created (in which several compounds appear more than once), and the regression line for each new set is calculated. This process is summarised in Table 1, and two of these bootstrap samples are depicted in Fig. 2.

By performing this resampling scheme many times, a good estimate can be obtained of the distribution of the statistics of interest, in this case slope and intercept. The distributions, obtained with 1999 resamplings, are shown in the histograms in Fig. 3. These distributions can be seen as approximations to the true distributions of the estimators, and therefore statistics of interest such as bias, standard deviation, and confidence intervals can be derived from them in the usual manner.

At first sight, such a procedure may resemble a fraud: the name "bootstrap" is a reference to the famous story of the Baron Von Münchhausen, who pulled himself up by his bootstraps out of a swamp. However, the basic idea is actually valid, although in many cases the "naive" bootstrap scheme outlined previously should be refined to be accurate. Its very simplicity makes it applicable in a large range of problems.

In the next sections, we will go into the basics and show why the bootstrap works. Several improvements over the naive bootstrap estimators for common situations are given, and are compared with al-

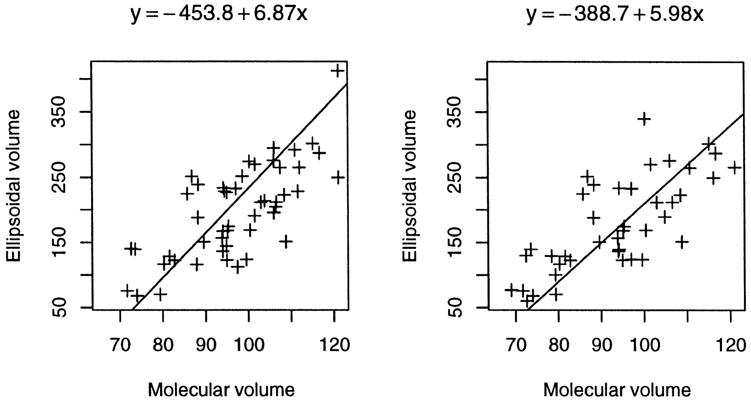

Fig. 2. The first two bootstrap samples from Table 1 and the corresponding orthogonal regression lines.

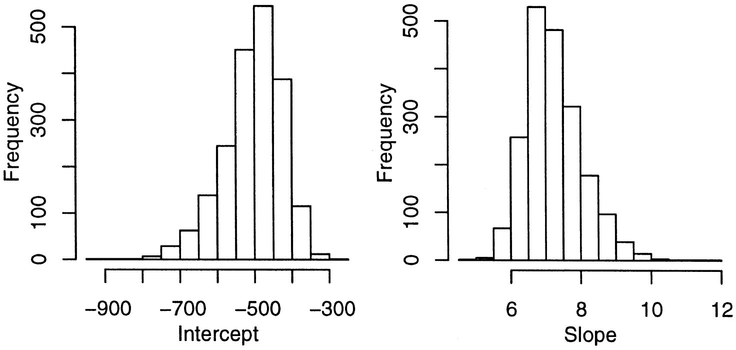

Fig. 3. Distribution of 1999 slopes and intercepts from bootstrap samples.

ternative approaches that are often more well known in the chemometrics community, such as cross-validation, jackknife and permutation tests. Practical issues concerning the application of bootstrap methods are addressed in a separate paragraph, and the paper concludes with a short review of some of the applications of bootstrap methods in chemometrics. First, however, it is necessary to define some notation.

2. Notation and principles

In this section, we briefly outline the notation used in this tutorial; it is mostly in agreement with the standard bootstrap literature. The cumulative distribution function of a random variable X, with observed values $x_1, x_2, \ldots, x_n$, is indicated by the symbol F. X may also be multivariate, and the observed value x_i then is a vector, e.g. $x_i = (MV_i, EV_i)$, the molecular and ellipsoidal volume of compound i. An unknown parameter θ, e.g. a correlation, or slope and intercept obtained from a regression, can be expressed by applying a function t to the distribution function:

$$\theta = t(F).$$

The symbol t is also used in relation to confidence intervals and hypothesis testing; from the context, it should be clear whether t is a function operating on the (empirical) distribution function, an entry from a t-table, or a test statistic.

Unfortunately, we usually do not know the complete distribution, but only have access to a random sample $\boldsymbol{x} = (x_1, \ldots x_n)$ from F, constituting the empirical distribution function $\hat{F}$. We therefore have to estimate θ from $\hat{F}$ or, using function u, from the data $\boldsymbol{x}$:

$$\hat{\theta} = t(\hat{F}) = u(\boldsymbol{x}).$$

This is called the plug-in principle, and it usually works quite well.

A confidence interval for the parameter θ is usually obtained from the estimator $\hat{\theta}$ by considering the probability distribution of $\hat{\theta} - \theta$. If s_α denotes the α-percentile of the distribution of $\hat{\theta} - \theta$, then a confidence interval for θ is based on the probability statement

$$P\left(s_{\alpha/2} \leq \hat{\theta} - \theta \leq s_{1-\alpha/2}\right) = 1 - \alpha,$$

which, after rewriting, leads to the interval

$$\hat{\theta} - s_{1-\alpha/2} \leq \theta \leq \hat{\theta} - s_{\alpha/2}.$$

The coverage probability, the probability of containing the true unknown value of θ, equals $1 - \alpha$. Note that, in order to be able to use this interval, it is necessary to know or to be able to approximate the distribution of $\hat{\theta} - \theta$.

It is more common to use the distribution of the studentised estimator $(\hat{\theta} - \theta)/\widehat{\mathrm{se}}$, where $\widehat{\mathrm{se}}$ is an estimator of the standard error of the estimator $\hat{\theta}$. This random variable often has an approximate t-distribution with $df = n - p$ degrees of freedom, where p is the total number of unknown parameters to be estimated from the data. Let $t_{df;\,\alpha}$ denote the α-percentile of the t-distribution with df degrees of freedom. Analogous to the above reasoning, the following well-known confidence interval can be obtained:

$$\hat{\theta} - t_{df;\alpha/2}\widehat{\mathrm{se}} \leq \theta \leq \hat{\theta} + t_{df;\alpha/2}\widehat{\mathrm{se}}.$$

Because the t-distribution is symmetric about 0, usually only the upper $\alpha/2$-quantile, $t_{df;\ \alpha/2}$ $(= -t_{df;\ 1-\alpha/2})$ of the t-distribution is used. If the normal approximation is used, then $t_{df;\ \alpha/2}$ may be replaced by $z_{\alpha/2}$. Both confidence intervals have approximate coverage probability $(1-\alpha)$ 100%; approximate because the distributions of $\hat{\theta} - \theta$ and its studentised version are approximated by a t-distribution and a standard normal distribution, respectively.

3. The bootstrap: general ideas

In general, we would like to obtain both the estimate $\hat{\theta}$ and some measures for its accuracy, such as bias and standard error. For many statistics, such as e.g. a mean, these can be calculated analytically, but in case of more complicated statistics there is usually no analytical formula available. In such cases, the bootstrap can be used. It is based on the application of the plug-in principle to so-called bootstrap data sets $\boldsymbol{x}^*$, which are drawn from a distribution close to the unknown underlying distribution F:

$$\hat{\theta}^* = t(\hat{F}^*) = u(\boldsymbol{x}^*).$$

Here the asterisk is the conventional way to indicate a bootstrap sample.

More specifically, we are interested in the distribution of our estimator $\hat{\theta}$ around the true value θ. The bootstrap is based on the idea that the variability of $\hat{\theta}$ around θ is mimicked by that of $\hat{\theta}^*$ around $\hat{\theta}$. This variability includes random variation (as measured, for instance, by the standard error) as well as systematic variation (bias). In some cases, it is possible to calculate the bootstrap estimators analytically, but in general resampling is necessary.

The bootstrap exists in a nonparametric and a parametric version. In the nonparametric setup, as already seen in the example from the Introduction, resampling is performed from the empirical distribution with replacement. This leads to a distribution of the statistic of interest, from which parameters such as standard errors can be estimated. In the parametric bootstrap, the underlying distribution F is estimated from the data by a parametric model, for instance a normal distribution. Bootstrap samples are then created by sampling from this model, rather than from the empirical distribution. Again, applying function t to each of these bootstrap samples leads to estimates $\hat{\theta}^*$, and from the distribution of estimated values, conclusions can be drawn. In practice, one of the main reasons for using bootstrap methods is uncertainty whether certain assumptions hold, and in most applications the nonparametric bootstrap is used. This will also be the main focus of the current paper.

Naive bootstrap estimates, based on the observed distribution of the statistic of interest such as the histograms of slope and intercept in Fig. 3, in most cases provide inaccurate answers. Many improvements have been suggested, and the most important ones will be discussed now.

4. Bootstrap point estimates

A discrimination will be made between point estimates, interval estimates and hypothesis testing applications. Point estimates, as the name suggests, are estimates consisting of one number, e.g. measures of bias and standard error. Interval estimates consist of an upper and lower bound, and are treated separately, although in some cases they may be derived from point estimates (such as a standard error).

4.1. Bias and standard error

Estimates such as slope and intercept are called point estimates. For the example in the Introduction, the bootstrap distributions for slope and intercept are used to estimate standard errors:

$$\widehat{\mathrm{se}}_B = \sqrt{\frac{1}{B-1}\sum_i \left(\hat{\theta}_i^* - \hat{\theta}_.^*\right)^2},$$

where B is the number of bootstrap samples, $\hat{\theta}_i^*$ the i-th bootstrap estimate, and $\hat{\theta}_.^*$ the mean value of the bootstrap estimates. For the example from the Introduction, $\hat{\theta}_.^*$ is (7.24, −504.0), leading to standard error estimates of 0.795 and 77.0 for slope and intercept, respectively.

Apart from the variation around a mean value, the accuracy of a point estimate is determined by bias, the deviation of the expected value of an estimator from the true value. A bootstrap estimate for bias can be obtained from

$$\widehat{\text{bias}}_B = \hat{\theta}^*_{\cdot} - \hat{\theta},$$

i.e. the difference between the mean of the bootstrap distribution of parameter θ^* and the estimate from the empirical distribution $\hat{\theta}$ (7.18, −497.3). For slope and intercept in our example, this leads to bias estimates of 0.072 and −7.12, respectively. Compared to the standard errors, the bias in both cases is reassuringly small. If it were too large to ignore, one could correct for the bias:

$$\hat{\theta}_{\text{bc}} = \hat{\theta} - \widehat{\text{bias}}_B = 2\hat{\theta} - \hat{\theta}^*_{\cdot}.$$

One might think that the mean of the bootstrap distribution is the unbiased estimate, but this is not the case. On the contrary: if $\hat{\theta}^*_{\cdot}$ is greater than $\hat{\theta}$, the bias-corrected estimate will be *smaller*. Bias correction is not necessarily a good thing: the variability in a bias-corrected estimate may be increased to such an extent that it is safer to use the uncorrected estimate. One example where bias correction is necessary is in the estimation of prediction errors (see below).

4.1.1. Related methods

Other methods are available for estimates of standard error and bias. The most well-known one is the *jackknife*. In this method, n new data sets are created by eliminating each object in turn from the data set. In analogy with the bootstrap, the statistic of interest is calculated for each new data set, yielding n values for $\hat{\theta}_{(i)}$, where $\hat{\theta}_{(i)}$ is the estimated value from the sample with observation i removed. From the results, estimates of bias and standard error can be obtained:

$$\widehat{\text{bias}}_{\text{jack}} = (n-1)\left(\hat{\theta}_{(\cdot)} - \hat{\theta}\right)$$

$$\widehat{\text{se}}_{\text{jack}} = \sqrt{\frac{n-1}{n}\Sigma\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2},$$

where $\hat{\theta}_{(\cdot)}$, is the mean of the n values for $\hat{\theta}_{(i)}$. The jackknife estimates for bias for slope and intercept are 0.062 and −6.24, respectively. These values are very close to the values found with the bootstrap approach earlier. The jackknife standard error estimates for slope and intercept estimates, with values of 0.79 and 76.1, respectively, are close to the bootstrap estimates as well. Note that the factors $n-1$ and $(n-1)/n$ in the jackknife formulas for bias and standard error compensate for the fact that jackknife samples are much more similar to the original data set than bootstrap samples.

The jackknife often works well, provided that the statistic under study does not change drastically upon small changes in the data. As an example of a statistic that is not “smooth” in this sense, consider the median. In Table 2, the concentrations of cephradine still present it in solution after complexation are gathered for 17 complexating compounds. These concentrations give an idea of the efficiency of the complexation.

The median of these values is 6.9 mM. In Fig. 4, histograms are shown for the jackknife and bootstrap estimates of the median. Clearly, the estimate for the median is hardly influenced when one sample is deleted from the data set; only three different jackknife values are found. The corresponding standard error for the jackknife is 0.501 mM, which is much smaller than the bootstrap estimate of 2.238 mM (100 bootstrap samples). It can be shown that the jackknife is not consistent for the median, whereas the bootstrap, considering data sets that are less similar to the original data, is consistent.

4.2. Prediction error

As already mentioned, prediction error estimation is one example where a correction for bias is useful. In the example of the complexation of organic compounds with cephradine, one of the goals is to be able to predict whether or not co-crystallisation will oc-

Table 2
Efficiencies of complexation with cephradine, for 17 complex-forming compounds. The efficiencies are expressed as the concentration (mM) of cephradine still present after complex formation (initial cephradine concentration: 30 mM)

1.3	6.6	15.0	2.7	3.4	2.9
8.6	10.3	16.4	18.1	6.9	5.5
16.0	3.7	14.4	1.1	21.8	

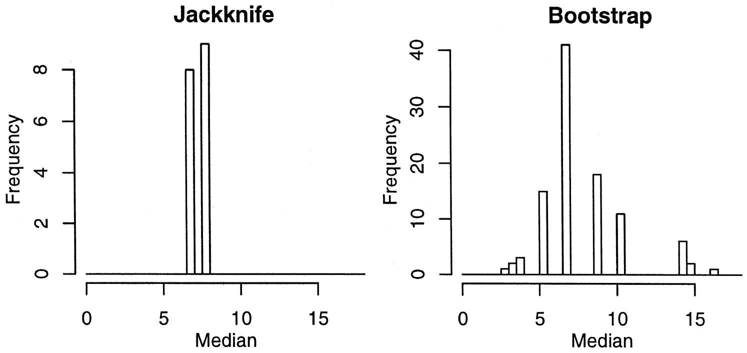

Fig. 4. Median values of 17 jackknife samples (left) and 100 bootstrap samples (right). The variability in the bootstrap samples is clearly much larger.

cur. Descriptors such as molecular volume in these cases are too general, and therefore similarities of the 120 compounds with respect to shape and charge distributions were calculated. This was done by aligning each pair of compounds with respect to an optimal shape overlap, and subsequently calculating the similarity in charge distributions. Both steps were performed using the program Tsar [8].

The resulting 120 × 120 matrix was used to construct PLS models. One to six latent variables were used. For six latent variables, only 7.5% of all compounds in the data set was predicted incorrectly. Obviously, this percentage, called the apparent error rate or the root mean square error of calibration (RMSEC), is not a good estimate of prediction error: it is biased (too optimistic) because the data are used both for the construction of the model and the error estimation.

Several bootstrap methods are available to obtain better prediction error estimates. The naive bootstrap estimator would create bootstrap training sets by resampling, build a model for each bootstrap sample, and test each model on the original, complete, data set. Unfortunately, this estimate is too optimistic as well. One approach to compensate for this bias is called the *.632 bootstrap* estimate, given by

$$\widehat{\mathrm{err}}^{.632} = 0.368\,\mathrm{RMSEC} + .632\,\hat{\epsilon}_0 .$$

For each bootstrap sample, a model is built and tested against only those objects *not* present in the sample; $\hat{\epsilon}_0$ is the average over all bootstrap samples of the prediction errors. The factor 0.632 is due to the fact that this is approximately the probability that an observation is present in a bootstrap sample. Practical and theoretical evidence suggests that this is a very reliable estimator [2,3].

Applying the .632 estimator (using 200 bootstrap samples) leads to estimates of prediction error for one to six latent variables (the dashed line in Fig. 5). Clearly, the addition of more latent variables does not lead to better models, and the minimum prediction error lies around 30%. The RMSEC estimate (solid line in Fig. 5) greatly underestimates the error for models with more than one latent variable.

4.2.1. Related methods

Cross-validation [9], the most often used method for error estimation in chemometrics, is closely related to the jackknife. Whereas in bootstrap methods new data sets are created by resampling with replacement from the original data, in cross-validation one creates new data sets by systematically removing objects from the data set, either in small groups or individually (k-fold cross-validation and leave-one-out cross-validation, respectively). The residuals for the observations that were left out, using the model built with the remaining observations, serve as a measure for the overall prediction error. In most cases, depending on the size of the data set, cross-validation requires fewer calculations than the bootstrap. Whereas the leave-one-out cross-validation is unbiased, it may suffer from a large variability, especially in smaller samples. Leaving out m samples at a time (where $m \times k = n$) reduces this variability, but introduces a bias. In Fig. 5, the results of leave-one-out and leave-three-out cross-validations for the pre-

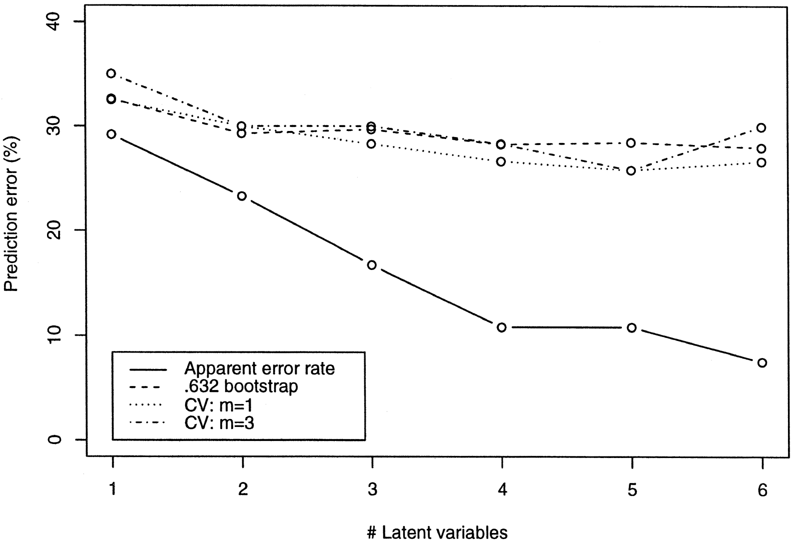

Fig. 5. Error percentages of predicting co-crystallisation behaviour with PLS for a varying number of latent variables. The apparent error rate (the error of training set) is too optimistic. Bootstrap and cross-validation approaches compare well.

diction of co-crystallisation using kernel-PLS are summarised. The error estimates are in good agreement with the .632 bootstrap.

5. Interval estimation

In general, a confidence interval for θ is more informative than a point estimate for θ alone, and several bootstrap methods have been proposed to obtain intervals with approximately correct coverage probabilities. In Section 2, we have reviewed classical approaches to constructing confidence intervals. The construction of confidence intervals is one of the areas where the bootstrap has achieved major success, and many different techniques are available (for a recent review in the area of medical statistics, see Ref. [10]). Here, we treat only the most popular variants.

In all of these, sample percentiles of the bootstrap distribution are used to estimate confidence intervals. The sample α-percentile of the bootstrap distribution is given by the $(B+1)\alpha$-th ordered value, where the ordering is such that $\hat{\theta}_1^* \leq \hat{\theta}_2^* \leq \ldots \leq \hat{\theta}_B^*$. For example, the $\alpha = 0.025$ and $\alpha = 0.975$ quantiles of a sample of size $B = 1999$, are the 50th and 1950th elements of the sorted bootstrap samples, respectively. In cases where α $(B+1)$ does not equal a whole number, interpolation can be used [3] (the somewhat odd-looking number of 1999 bootstrap samples is chosen to prevent the interpolation). The differences between the bootstrap methods arise from the choice of the statistic, or the way to convert percentiles to confidence intervals.

Basic bootstrap confidence intervals: The *basic* bootstrap confidence interval is based on the same simple basic rule that also underlies bootstrap estimation of bias and standard error: the distribution of $\hat{\theta} - \theta$ is approximated by that of $\hat{\theta}^* - \hat{\theta}$. This means that also the quantiles of the distribution of $\hat{\theta} - \theta$ are approximated by those of the bootstrap distribution of $\hat{\theta}^* - \hat{\theta}$. Thus, the percentiles of s_α of the distribution of $\hat{\theta} - \theta$ that are used in the confidence interval

$$\hat{\theta} - s_{1-\alpha/2} \leq \theta \leq \hat{\theta} - s_{\alpha/2}$$

of Section 2 are replaced by the appropriate quantiles of the bootstrap approximation, leading to the basic bootstrap confidence interval

$$\hat{\theta} - s^*_{((B+1)(1-\alpha/2))} \leq \theta \leq \hat{\theta} - s^*_{((B+1)/(\alpha/2))}.$$

It is easier, however, to look at the quantiles of $\hat{\theta}^*$ itself, instead of its centred version $\hat{\theta}^* - \hat{\theta}$. If r_α^* is the α-percentile of the distribution of $\hat{\theta}^*$ and s_α^* of $\hat{\theta}^* - \hat{\theta}$, then r_α^* is related to s_α^* by

$$r_\alpha^* = s_\alpha^* + \hat{\theta},$$

for the simple reason that the two distributions differ only by a shift $\hat{\theta}$. Inserting this relationship into the earlier obtained basic bootstrap interval gives

$$2\hat{\theta} - r^*_{((B+1)(1-\alpha/2))} \leq \theta \leq 2\hat{\theta} - r^*_{((B+1)(\alpha/2))}.$$

The 50th and 1950th sorted intercepts for the 1999 bootstrap samples in the introductory example are -688.0, and -375.6, respectively. This leads to a 95% confidence interval of $[-618.9, -306.5]$ for the intercept.

Percentile bootstrap confidence intervals: suppose that a transformation $u^* = h(\hat{\theta}^* - \hat{\theta})$ exists that makes the bootstrap distribution symmetric around 0. Then the symmetry relation $u^*_{\alpha/2} = -u^*_{1-\alpha/2}$ for the quantiles of u^* can be used, which, after back-transformation, leads to the confidence interval

$$r^*_{((B+1)\alpha/2)} \leq \theta \leq r^*_{((B+1)(1-\alpha/2))}.$$

This was the original approach of Efron [1], and these intervals are called the *percentile* bootstrap intervals. Remarkably, the symmetrising transformation does not have to be known to calculate these intervals. For the method to work well, it is required that such a transformation exists, however, and often this is not the case. In such a situation, the percentile method must be adjusted, and a well-known method to achieve this is the BC$_\alpha$ estimator [2]. In the example, the 95% confidence interval for the intercept is given by the 50th and 1950th sorted intercepts, which we have seen to be $[-688.0, -375.6]$.

Studentised bootstrap (bootstrap-t) confidence intervals: the coverage of both the basic bootstrap and the percentile bootstrap may be quite far off. One way to correct for this is to use *studentised* statistics (the bootstrap-*t* interval in Ref. [2]). In Section 2, we discussed how confidence intervals can be constructed from the studentised statistic $(\hat{\theta} - \theta)/\widehat{se}$ by approximating its distribution by either a *t*-distribution or a standard normal distribution. Another way is to approximate its distribution with the bootstrap. For each bootstrap sample $(b = 1, \ldots, B)$, calculate

$$t^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{\mathrm{se}}^*(b)},$$

where $\hat{\theta}^*(b)$ is the estimate from the *b*-th bootstrap sample and $\widehat{\mathrm{se}}^*(b)$ an estimate for the standard error from this sample. Sometimes an analytical expression for the standard error estimate is not available, as in the case of estimating the intercept using an orthogonal regression, and in such cases an inner bootstrap loop or a jackknife estimate may be used. Subsequently, $t^*_{((B+1)\alpha/2)}$ and $t^*_{((B+1)(1-\alpha/2))}$ are used as approximations of the quantiles $t_{\alpha/2}$ and $t_{1-\alpha/2}$ of the distribution of $(\hat{\theta} - \theta)/\widehat{\mathrm{se}}$. The studentised bootstrap confidence interval then becomes

$$\hat{\theta} - t^*_{((B+1)(1-\alpha/2))}\widehat{\mathrm{se}} \leq \theta \leq \hat{\theta} - t^*_{((B+1)\alpha/2)}\widehat{\mathrm{se}}.$$

The bootstrap-*t* confidence interval procedure works quite well, provided the standard errors are approximately independent of $\hat{\theta}^*(b)$. If this is not the case, a variance-stabilised bootstrap-*t* interval can be used [2,3].

In the example of the orthogonal regression lines, each bootstrap sample leads to an estimate of slope and intercept, plus standard error estimates for both (using an inner bootstrap loop with 50 bootstrap samples). These are used to calculate studentised statistics, as described above, which serve as a *t*-table, constructed especially for the data at hand. Several values are gathered in Table 3. Notice the mirrored asymmetry in the percentiles, as already suggested by Fig. 3.

Table 3
Percentiles of the normal distribution and the studentised bootstrap distributions for slope and intercept in the orthogonal regression case

Percentile	0.025	0.05	0.10	0.50	0.90	0.95	0.975
Normal $N(0, 1)$	-1.96	-1.64	-1.28	0.00	1.28	1.64	1.96
Studentised intercept	-1.69	-1.44	-1.13	0.03	1.39	1.86	2.26
Studentised slope	-2.30	-1.84	-1.42	-0.02	1.13	1.42	1.61

5.1. Discussion of bootstrap interval estimates

One feature of the bootstrap, which has been shown to be particularly helpful if the studentised bootstrap confidence intervals are used, is that it captures the skewness of the distribution of the statistic. Fig. 3 illustrates this point very well. In the classical approach, the distribution of $\hat{\theta} - \theta$ or $(\hat{\theta} - \theta)/\widehat{\text{se}}$ is approximated by for instance a normal distribution; a distribution that is symmetric. The bootstrap approximation correctly captures the skewness of $\hat{\theta}$ (Fig. 3) and uses it to obtain confidence intervals for θ, which are not symmetric around the point estimate $\hat{\theta}$. This leads to confidence intervals that have coverage probabilities that are closer to the desired coverage probability of $1 - \alpha$ than those based on the normal approximation [11].

In Fig. 6, the estimates for the basic, percentile and studentised confidence intervals are gathered for both slope and intercept. For both parameters, the studentised and basic intervals agree quite well, whereas the percentile bootstrap shows somewhat different intervals. This is caused by the fact that the percentile bootstrap confidence interval uses the skewness of the bootstrap approximation exactly the wrong way around. The percentile method should not be used for skew distributions.

The studentised interval estimates are generally considered to be very reliable, and have shown desirable properties, both in theory and in practice. An important disadvantage of studentised bootstrap confidence intervals is that they may be very time-consuming to calculate, notably in cases where there is no analytical expression for standard errors. The coverage probabilities of basic as well as percentile methods in general are less accurate, although basic bootstrap intervals are very useful in estimating confidence intervals for statistics like the sample median. Furthermore, both are attractively simple to implement.

Percentile intervals are regularly used, mainly because of two appealing properties (apart from their ease of implementation): they are transformation-respecting, and they always lead to valid intervals. The latter property means that an interval for a correlation coefficient (for example) will always be in the range $[-1, 1]$, which is not true for the other intervals. The transformation-respecting property means that a confidence interval for parameter θ will be equal to the confidence interval calculated for $g(\theta)$ after back-transformation: consider the correlation between molecular volume and ellipsoidal volume. Calculation of the 95% bootstrap percentile intervals for log-transformed data will lead to the same interval as taking the logarithm of the bootstrap interval obtained from the original data.

6. Hypothesis testing

The subject of hypothesis testing is intimately related to confidence intervals. With a hypothesis test, we investigate the evidence presented by the data against a null hypothesis H_0 of “no difference”. The usual advantages of bootstrap methods apply; one does not have to assume a specific distribution of the data, but can rely on the empirical distribution. Fur-

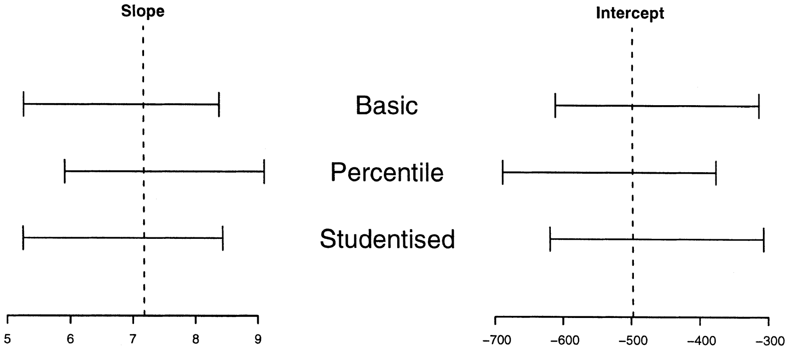

Fig. 6. Comparison of bootstrap confidence intervals ($B = 1999$) for the orthogonal regression line example. The estimated values for slope and intercept are indicated with the dashed line. The basic and studentised intervals agree well, with the latter being slightly narrower.

thermore, because of the plug-in principle, bootstrap methods are widely applicable.

An example can be seen in Fig. 1, where the molecular volume was plotted against ellipsoidal volume for a number of compounds. One may now wonder whether there is a difference between the orthogonal regression lines for complexating and non-complexating compounds. In particular, one might want to check whether the intercept is significantly different for the two lines, regardless of a difference in slope. The intercept for the complexating compounds is $a_c = -497.3$, and for the non-complexating compounds the intercept is $a_{nc} = -901.4$.

Because of the intimate connection between confidence intervals and hypothesis testing, it should come as no surprise that the use of studentised statistics increases the accuracy of the bootstrap with hypothesis testing as well. Therefore, we use the studentised statistic

$$t = \frac{a_c - a_{nc}}{\sqrt{s_{a_c}^2 + s_{a_{nc}}^2}},$$

where the use of $s_{a_c}^2$ and $s_{a_{nc}}^2$ indicates we do not assume the variances of both intercept to be equal. Note that t is of the form $(\hat{\theta} - \theta)/\widehat{\mathrm{se}(\theta)}$, with $\hat{\theta} = a_c - a_{nc}$, and $\theta = 0$ under H_0. Bootstrap samples are constructed with 68 data points, sampled with replacement from the complexating compounds, and 52 data points sampled with replacement from the non-complexating compounds. For each bootstrap sample, the corresponding orthogonal regression lines are calculated. Again, estimates for standard errors are obtained using an inner bootstrap loop with 50 bootstrap samples. For hypothesis testing, we wish to approximate the distribution of the test statistic under H_0. Under the null hypothesis, the expected value of the numerator $\hat{\theta} - \theta$ equals 0. In applying the bootstrap, we should therefore make sure that the same is true for the bootstrap statistic. This leads to the studentised bootstrap test-statistic

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{\widehat{\mathrm{se}}(\hat{\theta}^*)} = \frac{a_c^* - a_{nc}^* - (a_c - a_{nc})}{\sqrt{s_{a_c}^{*2} + s_{a_{nc}}^{*2}}},$$

where $\hat{\theta} = a_c - a_{nc}$, the observed difference in intercept.

In the left panel of Fig. 7, the differences in intercept $a_c^* - a_{nc}^*$ for 1999 bootstrap samples are depicted. The 1999 values for the studentised test-statistic are depicted in the right panel. The observed value of the test statistic, where the standard errors are obtained from 50 bootstrap samples,

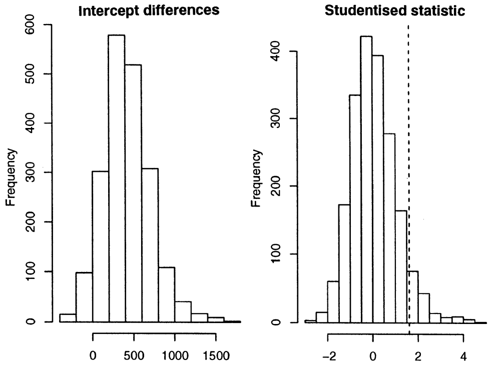

Fig. 7. Differences in intercept between complexating and non-complexating compounds (left-panel). A histogram of the studentised test stastistics is shown on the right, with $t_0 = 1.62$ indicated by the dashed line.

$$t_0 = \frac{a_c - a_{nc}}{\sqrt{s_{a_c}^2 + s_{a_{nc}}^2}} = 1.62,$$

is indicated by the dashed line. The p-value for the hypothesis test is given by the fraction of bootstrap results, larger than the observed value for the test statistic:

$$p = \frac{1 + \#(t^* \geq t_0)}{B + 1},$$

which in this case is $131/2000 = 0.066$. Although there is borderline evidence to reject the null hypothesis, we would not reject it at the 5% level. Similarly, the slopes are not significantly different at this confidence level ($p = 0.068$).

6.1. Related methods

An alternative to bootstrap methods in hypothesis testing is sometimes formed by *permutation tests*. In a two-sample setting, a permutation test tests the null hypothesis that there is no difference between two distributions F and G:

$$H_0 : F = G.$$

The test is performed by repeatedly and randomly assigning observations to one of the two distributions, and for each permutation calculating a test statistic, such as a difference in means. The distribution of the test statistic θ^* is then compared to the observed difference in means, θ, for the two samples. The evidence against the null hypothesis is given by the fraction of θ^*, larger than θ. One might, for instance, reject the null hypothesis if less than 5% of the values in θ^* are larger than θ. Other test statistics may be used, too. As an application example, van der Voet [12] describes an application where permutation tests are used to assess the predictive accuracy of different models.

Applying a simple permutation test (1999 permutations) to the intercepts of complexating and non-complexating compounds yields a p-value of 0.109, so the null hypothesis will not be rejected at the 0.05 level. Note that the null hypothesis in this test is stricter than the bootstrap hypothesis test, where we did not assume slopes to be equal. A bootstrap test, analogous to the permutation test in the example, would resample from the complete set of 120 objects, assign the first 68 compounds to the complexating class and the remainder to the non-complexating class, and calculate the difference in intercepts. The only difference with the permutation test would be that resampling is done with replacement. In the bootstrap hypothesis test described earlier, resampling was performed separately within the two classes (stratified resampling).

Both the permutation test and bootstrap approach have their advantages. The permutation test is exact; the significance level found by the test is exactly equal to the chance of finding a test statistic as extreme as the one observed. The bootstrap is not exact, but usually gives quite a good estimate. The main advantage of the bootstrap is that it can be applied more widely. In our example, for instance, we tested $a_c = a_{nc}$ without believing that the lines of the complexating and non-complexating compounds have the same slopes, which is impossible to do with permutation tests.

7. Practical issues

In this section, we discuss a number of practical issues. First we go into the important practical problem of determining how many bootstrap samples to take in a particular application of the bootstrap. We go on to discuss two different ways of using the bootstrap in linear regression. The last part of this section lists a number of adaptations of the bootstrap that are useful in particular situations.

7.1. On the number of bootstrap samples

In some isolated instances it is possible to calculate bootstrap estimators analytically, but typically Monte Carlo simulation is needed to approximate the bootstrap estimator. The procedure was described in Section 3; here the aim is to give some guidelines concerning the number of bootstrap samples to take in the Monte Carlo (B).

Most authors recommend to take 100 to 500 bootstrap samples. With ever increasing computing power, there is hardly ever an objection to go beyond that number. In considering the number of bootstrap samples to take, it is worth having in mind that the error of a Monte Carlo bootstrap approximation to the distribution of a statistic is the sum of two independent errors from different sources. The first is what we will call the bootstrap or statistical error; the second is the Monte Carlo or simulation error. The bootstrap error is unavoidable and is independent of B; by altering B we can only influence the Monte Carlo error and we should try to choose B in such a way that the Monte Carlo error is no larger (and preferably a lot smaller) than the bootstrap error.

To give an impression of the Monte Carlo error and how it varies with B, we conducted a simulation study. The “true” underlying model is the one that is fitted in the orthogonal regression in Section 1. The “true” standard errors of the estimates of intercept and slope were taken to be the values that were found in Section 4 (77.0 and 0.795, respectively). We generated one sample from this model and calculated the bootstrap standard error of the estimates of slope and intercept, again with a Monte Carlo simulation using 100,000 bootstrap samples. This resulted in values of 83.31 and 0.836 for intercept and slope, respectively. The idea is that $B = 100{,}000$ is large enough to be able to neglect the Monte Carlo error. Thus, we call the bootstrap standard error with $B = 100{,}000$ the “theoretical bootstrap standard error”, and we think of the difference between true standard error and theoretical bootstrap standard error from this one sample as caused by only the statistical error. To assess the Monte Carlo error, we used the same sample and calculated the bootstrap standard error repeatedly for smaller values of B. Histograms of the resulting bootstrap standard errors are shown in Fig. 8 for the intercept. True standard error and theoretical bootstrap standard error are indicated in each histogram. It is clearly seen that for small values of B, the statistical error is dwarfed by the simulation error. For $B = 1000$, the Monte Carlo error is acceptably small.

This has two implications. The first is that the more observations we have in our dataset, the more accurate the theoretical bootstrap approximation is and the higher B should be taken in order to make the Monte Carlo error accordingly small. The opposite is what is usually done in practice to control computing time. Unless computing resources are severely restricted, we recommend to increase, not decrease B with growing sample size.

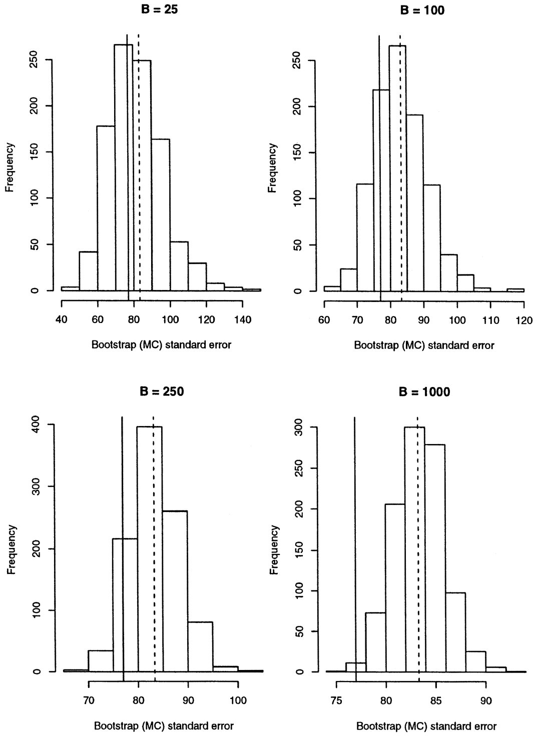

Fig. 8. Histograms of 1000 simulated bootstrap standard errors for increasing values of B. The true standard error and theoretical bootstrap standard error are indicated in each histogram by straight line and dashed line, respectively. The variability in the bootstrap estimates decreases with increasing B.

The second implication is that a higher value of B is needed for more refined bootstrap estimates that are intended to increase the accuracy of the bootstrap approximation, such as studentised bootstrap estimates. If it is decided that there is merit in attempting to increase the accuracy of the bootstrap approximation, then it would be unwise to lose this advantage by choosing too low a value of B, resulting in a Monte Carlo error that is too large in comparison to the bootstrap error. The point of the first implication, that a higher value of B is needed for larger samples, is even more important here. To assess the contribution of the bootstrap error and Monte Carlo error, a method called jackknife-after-bootstrap [13] can be used. Davison and Hinkley [3] give the practical rule of thumb of taking $B \approx 40\ n$, where n is the sample size.

We conclude this subsection by noting that improved ways of resampling, like balanced sampling or importance sampling (see Ref. [3] and Section 7.3 below) can be used to obtain the same Monte Carlo error with a lower number of bootstrap samples.

7.2. Regression: resampling cases or residuals

There are two fundamentally different approaches in applying the bootstrap to linear regression problems. We will describe each approach in a simple univariate linear regression setting, where $(x_1, y_1) \ldots (x_n, y_n)$ are observed such that

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j.$$

Here $\varepsilon_1 \ldots \varepsilon_n$ are independent measurement errors with mean zero. Extension to multivariate linear regression is straightforward. After describing the two approaches, we shall discuss which approach to take, depending on the context.

Resampling cases: A new bootstrap dataset is created by sampling independently from the pairs $(x_1, y_1), \ldots, (x_n, y_n)$, yielding $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$. Simulated values $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are then computed from $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$ in the same way that $\hat{\beta}_0$ and $\hat{\beta}_1$ were computed from the original data. This is the approach taken in the example from the introduction.

Resampling residuals: The residuals are used to make new bootstrap datasets. More precisely, suppose $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of β_0 and β_1 and e_1

$= y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1$, ..., $e_n = y_n - \hat{\beta}_0 - \hat{\beta}_1 x_1$ are the residuals. Then a new dataset is made by sampling independently from e_1, ..., e_n, yielding, e_1^*, ..., e_n^*, and by setting

$$y_j^* = \hat{\beta}_0 + \hat{\beta}_1 x_j + e_j^* .$$

Simulated values $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ of the coefficient estimates may then be computed from the bootstrap data (x_1, y_1^*), ..., (x_n, y_n^*) in the same way that $\hat{\beta}_0$ and $\hat{\beta}_1$ were computed from the original data, e.g. by least squares. It is also possible to do parametric resampling on the basis of the residuals. If it is assumed, for instance, that the measurement errors are normally distributed with mean zero and unknown variance σ^2, then σ^2 can be estimated from the residuals by $\hat{\sigma}^2 = \Sigma(e_j - e.)^2/(n-2)$. Then, resampled errors e_1^*, ..., e_n^* can be created by generating mean zero random variables with variance $\hat{\sigma}^2$. Simulated values of the coefficient estimates are obtained as in nonparametric resampling.

The bootstrapping residuals method is the method to take if the independent variables (covariates) x_j are controlled by design. The method is very sensitive to the assumption that the errors are stochastically independent, and that their distribution is independent of the covariates. If the variance of the errors depends on the covariates, an adaptation, called the wild bootstrap, exists which can be used instead (see Section 7.3 below). Alternatively, it is possible to resample from the pairs instead of the residuals.

If the covariates are also random, then a good bootstrap procedure should mimic the bivariate distribution (x_1, y_1), ..., (x_n, y_n). This is exactly what is done when cases or pairs are resampled. Resampling cases is safer to use in general: it will also work if the assumptions necessary to use residual resampling do hold.

7.3. The bootstrap: variations on a theme

A wide variety of adaptations of the bootstrap, carrying colourful names, have been proposed over the years, each tailored to a specific application or goal. It would go too far to mention them all; here is a biased list with a short description.

Wild bootstrap: in a regression setting with heteroscedasticity (i.e. the standard deviation of the errors varies with different values of the covariates), resampling from the residuals gives inconsistency because in the resampling paradigm it is tacitly assumed that residuals form an independent sample from a common distribution. The wild bootstrap proposes to multiply each residual independently with a random variable with expectation zero and variance one. More information can be found in Section 6.2.6 of Ref. [3].

Smoothed bootstrap: in some applications it is advantageous not to resample from the empirical distribution but from a smoothed empirical distribution. This method improves the nonparametric bootstrap in some non-smooth cases, like the median. The smoothed bootstrap falls in a category of bootstrap methods that obtain bootstrap samples by sampling from an estimate of the underlying distribution other than the empirical. Such an estimate of the underlying distribution is then termed a *resampling distribution*. Many instances of this approach are scattered along the literature. In cases where the nonparametric bootstrap fails (see Section 8) many researchers have come up with a resampling distribution that worked. In fact, one can always find such a resampling distribution [14].

m-Out-of-n-bootstrap: a very general and promising way of resolving bootstrap failure, consisting of forming smaller bootstrap samples. Research is still active in this field. An example of the *m*-out-of-*n* bootstrap is given in Section 8.

Double (iterated) bootstrap: based on the principle of resampling from bootstrap samples. It is used both as a diagnostic tool (variation in bootstrap estimates is indicative of bootstrap failure) and as a calibrating tool. There the iterated bootstrap is used to estimate the departure of bootstrap coverage probability from the nominal level of confidence. The confidence level used in the bootstrap can be adjusted on the basis of the estimated departure. Very time-consuming (for the computer).

Balanced bootstrap: the balanced bootstrap is a method inherited from Monte Carlo methods developed before the bootstrap was known. It is a way to improve the Monte Carlo approximation for a given number of bootstrap samples by balancing the occurrence of the sample elements in the super-bootstrap sample (the sample of size $n \times B$ that is obtained by concatenating the B bootstrap samples). The interested reader is referred to Ref. [15] for more information.

Blocked bootstrap: in case of dependent observations, the ordinary bootstrap fails as well, since bootstrap samples are drawn independently from the original sample. A way to overcome this is to resample blocks of consecutive observations, big enough to retain the dependence present in the original sample, small enough to have enough blocks to sample from. Two versions are used; one is merely dividing the original sample of size n into k disjoint blocks of size m, such that $n = km$. A bootstrap sample is then obtained by sampling independently from the k blocks and concatenating the k blocks of size m into one bootstrap sample of size n. The second version is called the *moving block bootstrap*. It divides the original sample into $n - m + 1$ overlapping blocks of consecutive sample elements of size m. The first block contains $x_1, \dots, x_m$, the second $x_2, \dots, x_{m+1}$, etc. A bootstrap sample is now obtained by forming an independent sample of size $k = n/m$ from the $n - m + 1$ overlapping blocks and again concatenating the k blocks of size m into one bootstrap sample of size n.

8. To pull yourself up... or not

It should not be imagined that the bootstrap will always give the results we want. The preceding sections have already shown that in many cases adaptations are needed to allow a proper implementation of the bootstrap in a statistical analysis. The adaptations we have discussed so far are merely refinements in the sense that they improve on an already moderately (but not entirely) successful application of the bootstrap. In some situations, the bootstrap can give downright disastrous results, which are nowhere near the correct answer.

Consider the data of Table 2, and suppose we want to estimate a “maximum attainable complexating efficiency”. That would correspond to the smallest possible rest concentration in a dataset like that in Table 2. We view the data $x_1, \dots, x_n$ as a random sample, taken from a distribution F whose support is bounded on the left by the unknown parameter (smallest concentration possible) θ. The most obvious estimator of θ is the maximum likelihood estimator $\hat{\theta} = \min(x_1, \dots, x_n)$. This estimator, however, is biased, because $\hat{\theta}$ will always be less than or equal to θ.

We could try to use the bootstrap to estimate the bias of our estimator and obtain a bias-adjusted estimator, or we could apply the bootstrap to estimate the standard error of $\hat{\theta}$ or to construct a confidence interval for θ. The statistic $\hat{\theta}$ is non-smooth in the same sense as the median, discussed in Section 4, so we would expect the jackknife to give erroneous results. It appears that $\hat{\theta}$ is non-smooth to such a degree that even the bootstrap estimate of standard error is inconsistent. Fig. 9 shows a histogram of 1000 bootstrap values of the standardised statistic $n(\hat{\theta} - \theta)$ using the nonparametric bootstrap and a parametric bootstrap, assuming an underlying exponential distribution. It is apparent from the histogram that the nonparametric bootstrap is too discrete. In fact, it is easy to show that, for example, the probability $P(\hat{\theta}^* = \hat{\theta})$ is equal to $1 - (1 - 1/n)^n \approx 0.632$ for moderate to large values of n. For the parametric bootstrap, this probability is zero. Clearly, estimated quantiles based on the nonparametric bootstrap will be nonsensical.

What adaptation is possible to make the bootstrap work? One of them is the m-out-of-n bootstrap. The m-out-of-n bootstrap uses a smaller sample size m in the bootstrap statistic. Thus, $\hat{\theta}_m^*$ is now the minimum of a sample $x_1^*, \dots, x_m^*$ of size m, independently drawn from the full sample $x_1, \dots, x_n$. The subscript m in $\hat{\theta}_m^*$ is used to denote the fact that the bootstrap statistic is based on a bootstrap sample of

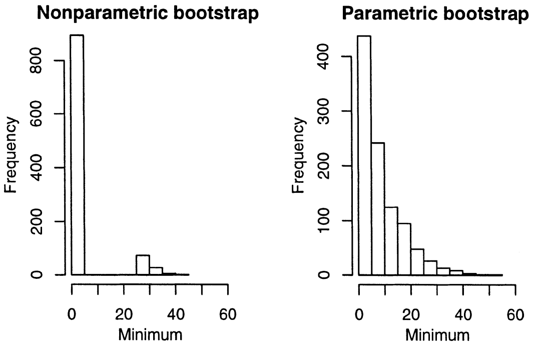

Fig. 9. Histogram of 1000 bootstrap values from the nonparametric bootstrap (left) and a parametric bootstrap (right), assuming an underlying exponential distribution. The histogram on the right suggests a much smoother distribution.

size m. The distribution of $m(\hat{\theta}_m^* - \hat{\theta}_n)$ is used to approximate that of $n(\hat{\theta}_n - \theta)$. The normalisation in these statistics is important and is such that the distribution of $n(\hat{\theta}_n - \theta)$ converges to a limit. The reason that the nonparametric bootstrap fails is that the empirical distribution is not a good estimate of the underlying distribution F in the extreme tails. Taking a small sample of size m from the empirical distribution based on n observations is like taking a sample of size n from the empirical based on N (more than n) observations, such that the ratio m/n remains approximately n/N. The empirical based on N observations is much closer than the empirical based on n observation and does succeed in estimating F in the tails. The bootstrap distribution of $m(\hat{\theta}_m^* - \hat{\theta}_n)$ actually estimates that of $m(\hat{\theta}_m - \theta)$, but since the statistic of interest is normalised so that it converges to a limit, there is not much difference between the distributions of $m(\hat{\theta}_m - \theta)$ and $n(\hat{\theta}_n - \theta)$, at least not for large m and n. The idea is as miraculous as other ideas we have seen in this tutorial, and again it works, not only for this example, but for almost any situation with independent observations from a common distribution where the statistic of interest can be normalised in such a way that it converges to a limiting distribution, cf. Ref. [16]. An important drawback is that you need quite a large dataset to effectively use the m-out-of-n bootstrap, large enough that the smaller sample m is large. Another is that it leaves a decision as to which resampling size m to take. Nevertheless, we apply the m-out-of-n bootstrap to our 17 data points. Histograms of 1000 bootstrap samples are shown for $m = 5$ and $m = 10$ in Fig. 10. The m-out-of-n bootstrap with $m = 5$ seems to perform best, despite the very small resampling size.

There are other ways of resolving the problem of inconsistency of the nonparametric bootstrap. One of them is using a parametric bootstrap, as already shown in Fig. 9, but that leaves the question which parametric model to choose. A nonparametric solution was suggested in Ref. [17]; the nonparametric estimator proposed there consists of using the empirical distribution and approximating it by a uniform distribution in the tails. The proposed estimator of the underlying distribution is very particular to this statistic; in principle, any new estimation problem might require a new resampling distribution. Fig. 10 shows a histogram of 1000 bootstrap values of $\hat{\theta}_n^*$ using this approach.

Now, let us go back to our problem of estimating the minimal obtainable rest concentration in the cocrystallisation example. Our initial estimator $\hat{\theta} = \min(x_1, \ldots, x_n)$ was based. The nonparametric bootstrap estimate of bias (which should not be trusted!) based on the Monte Carlo simulation of Fig. 9 equals 0.238; the parametric bootstrap estimates equals 0.507. The m-out-of-n bootstrap estimate with $m = 5$ and $m = 10$ and the alternative bootstrap estimate of Fig. 10 equal 0.447, 0.329 and 0.664, re-

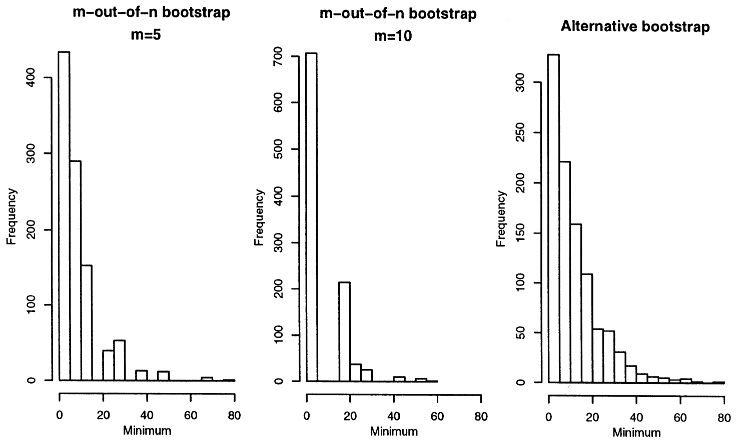

Fig. 10. Histogram of 1000 bootstrap samples from the m-out-of-n bootstrap (two left panels, $m = 5$ and $m = 10$, respectively) and an alternative nonparametric bootstrap [17].

spectively. Due to the small sample size, these estimates vary considerably, but all are closer to the parametric estimate than the original nonparametric approach. Taking the parametric bootstrap estimate of bias, the original estimate of 1.1 is adjusted to 0.59.

9. Conclusion

Bootstrap methods can be applied in a large number of statistical analyses, often providing answers of equal or higher quality than alternative methods. The price to be paid is computer time, which has become so cheap that this is hardly a serious objection. Why, then, is the bootstrap still not a standard tool in the chemometrical world? One of the reasons probably is the large number of different bootstrap methods, each with particular strengths and weaknesses, which are not always clear to the non-statistician. In this tutorial, our objective was to highlight several of the more popular bootstrap methods for common situations in chemical data analysis, and to compare the different variants with more commonly used techniques. It should provide the interested reader a platform of understanding from which to depart in reading the more specialised bootstrap literature.

A second reason for the less-than-abundant use of bootstrapping in the chemometrics may be the lack of available software. Although several high-quality packages exist (most notably the "boot" library for the S language [18], accompanying Ref. [3][1]), in most statistical packages only limited bootstrap functions are present, if at all. In Ref. [10] a short list can be found. However, it is not difficult to write a bootstrap library if none is available.

Although relatively small in number, the chemometrical applications of bootstrap methods cover a wide range of problems. Osten compares the distribution of the number of latent variables in PLS regression, selected according to several criteria with bootstrap methods [19]. Ortiz et al. [20] and Wehrens and van der Linden [21] apply bootstrap methods directly to assess the optimal number of latent variables in multivariate regression. The latter paper also discusses the use of bootstrap confidence intervals for elimination of uninformative variables, an approach similar to the jackknife procedure described in Ref. [22]. Whereas these bootstrap applications used the "resampling cases" scenario, several papers describe the use of "resampling residuals" in a regression context [23–25]. In Ref. [25], a naive bootstrap estimator performed quite badly in estimating prediction intervals in PLS regression. Several papers have appeared in which bootstrap methods are used in the validation of neural networks (e.g. Ref. [26]) or other classification methods, such as Classification and Regression Trees [27]. The methodology there is exactly the same; usually the preferred strategy is resampling cases.

An example of bootstrap methods for parametric hypothesis testing is described by Coakley and Simons [28], with the aim to detect inhomogeneities in isotope ratios, using both simulated and real data. Lodder and Hieftje [29,30] construct nonparametric confidence intervals to detect outlying samples in NIR analysis, an application intimately related to hypothesis testing. In the field of quantitative structure-activity relations, the bootstrap has been used to assess the variability of cluster parameters. Active compounds were clustered together, and the cluster was described using an ellipsoid, where confidence intervals for the ellipse parameters were obtained with a bootstrap approach [31]. Finally, Meinrath [32] recently described an application of the moving blocks bootstrap in UV–Vis spectroscopy.

Clearly, bootstrap methods offer many advantages in the analysis of real-world data, where we never can be sure about the validity of our assumptions. The most important drawback, the increase in computer time, is likely to become less and less serious. We hope this tutorial will contribute to a better understanding and a more widespread use of bootstrap methods in the chemometrics community, both at the level of application and of academic research.

Acknowledgements

Rene de Gelder and Ger Jan Kemperman are acknowledged for the data and stimulating discussions.

[1] Common implementations of the S language are S-Plus http://www.mathsoft.com/splus) and R http://cran.r-project.org); the examples in this tutorial are performed with R/boot combination.

Uwe Thissen is thanked for his comments on an earlier version of the manuscript.

References

[1] B. Efron, Bootstrap methods: another look at the jackknife, Ann. Stat. 7 (1979) 1–26.

[2] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[3] A.C. Davison, D.V. Hinkley, Bootstrap Methods and Their Applications, Cambridge Univ. Press, Cambridge, 1997.

[4] G.A. Young, Bootstrap: more than a stab in the dark? Stat. Sci. 9 (1994) 382–415, including discussion.

[5] G.J. Kemperman, R. de Gelder, F.J. Dommerholt, P.C. Raemakers-Franken, A.J.H. Klunder, B. Zwanenburg, Clathrate type complexation of cephalosporins with β-naphtol, Chem. —Eur. J. 5 (1999) 1205–1210.

[6] R. Wehrens, R. de Gelder, G.J. Kemperman, B. Zwanenburg, L.M.C. Buydens, Molecular challenges in modern chemometrics, Anal. Chim. Acta 400 (1999) 413–424.

[7] J.E. Jackson, A User's Guide To Principal Components, Wiley, New York, 1991.

[8] Tsar version 3.2, Oxford Molecular Group, Oxford, UK.

[9] M. Stone, Cross-validation choice and assessment of statistical predictions, J. R. Stat., Soc. B 36 (1974) 111–147.

[10] J. Carpenter, J. Bithell, Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, Stat. Med. 19 (2000) 1141–1164.

[11] H. Putter, W.R. van Zwet, Empirical Edgeworth expansions for symmetric statistics, Ann. Stat. 26 (1998) 1540–1569.

[12] H. van der Voet, Comparing the predictive accuracy of models using a simple randomization test, Chemom. Intell. Lab. Syst. 25 (1994) 313–323.

[13] B. Efron, Jackknife-after-bootstrap standard errors and influence functions, J. R. Stat., Soc. B 54 (1992) 83–127.

[14] H. Putter, W.R. van Zwet, Resampling: consistency of resampling estimators, Ann. Stat. 24 (1997) 2297–2318.

[15] P. Hall, The Bootstrap and Edgeworth Expansion, Springer, New York, 1992.

[16] P.J. Bickel, F. Götze, W.R. van Zwet, Resampling fewer than *n* observations: gains, losses, and remedies for losses, Stat. Sin. 7 (1997) 1–31.

[17] H. Putter, W.R. van Zwet, On a set of the first category, in: D. Pollard, E. Torgersen, G.L. Yang (Eds.), Festschrift for Lucien Le Cam, Springer, New York, 1997, pp. 315–324.

[18] R.A. Becker, J.M. Chambers, A.R. Wilks, The New S Language, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1988.

[19] D.W. Osten, Selection of optimal regression models via cross-validation, J. Chemometr. 2 (1988) 39–48.

[20] M.C. Ortiz, J. Arcos, L. Sarabia, Using continuum regression for quantitative analysis with overlapping signals obtained by differential pulse polarography, Chemom. Intell. Lab. Syst. 34 (1996) 245–262.

[21] R. Wehrens, W.E. van der Linden, Bootstraping principal-component regression models, J. Chemom. 11 (2) (1997) 157–171.

[22] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uniformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.

[23] P.L. Bonate, Approximate confidence intervals in calibration using the bootstrap, Anal. Chem. 65 (1993) 1367–1372.

[24] T. Roy, Bootstrap accuracy for non-linear regression models, J. Chemom. 8 (1994) 37–44.

[25] M.C. Denham, Prediction intervals in partial least squares, J. Chemom. 11 (1997) 39–52.

[26] E.P.P.A. Derks, L.M.C. Buydens, Aspects of network training and validation on noisy data: Part 2. Validation aspects, Chemom. Intell. Lab. Syst. 41 (1998) 185–193.

[27] S.L. Crawford, Extensions to the CART algorithm, Int. J. Man-Mach. Stud. 31 (1989) 197–217.

[28] K.J. Coakley, D.S. Simons, Detection and quantification of isotopic ratio inhomogeneity, Chemom. Intell. Lab. Syst. 41 (1998) 209–220.

[29] R.A. Lodder, G.M. Hieftje, Detection of subpopulations in Near-Infrared reflectance analysis, Appl. Spectrosc. 42 (8) (1988) 1500.

[30] R.A. Lodder, G.M. Hieftje, Quantile BEAST attacks the false-sample problem in NIR analysis, Appl. Spectrosc. 42 (8) (1988).

[31] A.P. Worth, M.T.D. Cronin, Embedded cluster modelling—a novel method for analysing embedded data sets, Quant. Struct.-Act. Relat. 18 (1999) 229–235.

[32] G. Meinrath, Robust spectral analysis by moving block bootstrap designs, Anal. Chim. Acta 415 (2000) 105–115.