Chapter 5

Bootstrap

5.1 Introduction

The bootstrap, proposed by Efron (1979), is by now considered as a breakthrough in statistics. Essentially, the bootstrap can be described as "simulating from an estimated model". This turns out to be tremendously useful for making statistical inference (confidence intervals and testing) and, analogous to the goals in cross-validation, for estimating the predictive power of a model or algorithm and hence also for tuning of statistical procedures.

5.2 Efron's nonparametric bootstrap

Consider the situation where the data are realizations of

$$Z_1, \ldots, Z_n \text{ i.i.d. } \sim P,$$

where P denotes an unknown distribution. The variables Z_i can be real-valued (usually then denoted by X_i), or they can be vector-valued. For example, $Z_i = (X_i, Y_i)$ are the pairs in a regression problem with $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$.

We denote a statistical procedure or estimator by

$$\hat{\theta}_n = g(Z_1, \ldots, Z_n) \tag{5.1}$$

which is a (known) function g of the data $Z_1, \ldots, Z_n$. The estimator $\hat{\theta}_n$ can be a parameter estimator or also a curve estimator (e.g. in nonparametric regression).

Whenever we want to make statistical inference, we would like to know the probability distribution of $\hat{\theta}_n$. For example, constructing a confidence interval for a true parameter θ requires the knowledge of the distribution of $\hat{\theta}_n$; or constructing a statistical test requires the distribution of $\hat{\theta}_n$ under the null-hypothesis. We also considered in chapter 4 the problem of estimating the generalization error in regression

$$\mathbb{E}[(Y_{new} - \hat{m}(X_{new}))^2]$$

which can be thought of as the expected value, a summary statistic of the distribution, of

$$\hat{\theta}_{n+1} = g(Z_1, \ldots, Z_n, Z_{new}) = (Y_{new} - \hat{m}_{Z_1,\ldots,Z_n}(X_{new}))^2, \quad Z_i = (X_i, Y_i),$$

where we have a function g of the training and test data.

Deriving the exact distribution of $\hat{\theta}_n$ is typically impossible, unless the function g is simple and P is a mathematically convenient distribution, e.g. P is a Normal distribution. If exact distributions are not available, much mathematical theory has been developed to get at least the asymptotic distribution of $\hat{\theta}_n$ as n gets large. For example, due to the central limit theorem,

$$\hat{\theta}_n = n^{-1} \sum_{i=1}^{n} X_i \approx \mathcal{N}(\mu, \sigma^2/n), \;\; X_i \in \mathbb{R},$$

where $\mu = \mathbb{E}[X_i], \; \sigma^2 = \mathrm{Var}(X_i)$. We then only need to estimate the parameters μ and σ in order to have an approximate distribution for $n^{-1} \sum_{i=1}^{n} X_i$. For the maximum-likelihood estimator in general, estimation of the asymptotic variance is already more subtle. Or for the sample median,

$$\begin{aligned}&\hat{\theta}_n = \mathrm{median}(X_1, \ldots, X_n) \approx \mathcal{N}(\theta, \sigma^2_{asy}/n), \;\; X_i \in \mathbb{R},\\ &\sigma^2_{asy} = \mathbb{E}[(X-\theta)^2]/(4f^2(\theta))\end{aligned}$$

where the asymptotic variance already involves quantities like the density f of P at the unknown parameter $\theta =$ median(P). Estimating this asymptotic variance is already a pretty awkward task, and we should keep in mind that we would then only get the asymptotic answer to the problem of getting the distribution of $\hat{\theta}_n$. Finally, for more complex algorithms, mathematical theory is lacking for obtaining the approximate, asymptotic distribution.

A pioneering step has then be taken by Efron (1979). Suppose we would know what the distribution P is: we could then **simulate** to obtain the distribution of any $\hat{\theta}_n$ with arbitrary accuracy (when simulating enough). Because we do not know the distribution P of the data generating mechanism, we use the **empirical distribution** $\hat{P}_n$ which places probability mass $1/n$ on every data point $Z_i, \; i = 1, \ldots, n$. The recipe is then to simulate from $\hat{P}_n$: generate simulated data

$$Z_1^*, \ldots Z_n^* \;\text{ i.i.d. } \sim \hat{P}_n.$$

Such a simulated new data set is called a bootstrap sample. We can now compute our estimator $\hat{\theta}_n^* = g(Z_1^*, \ldots, Z_n^*)$, analogously to (5.1) but based on the bootstrap sample, and we then repeat this many times to get an approximate distribution, e.g. via the histogram of many simulated $\hat{\theta}_n^*$'s.

5.2.1 The bootstrap algorithm

Bootstrapping an estimator as in (5.1) can be done as follows.

1. Generate a bootstrap sample

   $$Z_1^*, \ldots, Z_n^* \;\text{ i.i.d. } \sim \hat{P}_n.$$

   This can be realized as follows. Do n **uniform random drawings with replacement** from the data set $\{Z_1, \ldots, Z_n\}$, yielding the bootstrap sample.

2. Compute the bootstrapped estimator

   $$\hat{\theta}_n^* = g(Z_1^*, \ldots, Z_n^*),$$

   based on the bootstrap sample; the function $g(\cdot)$ is as in (5.1).

3. Repeat steps 1 and 2 B times to obtain

$$\hat{\theta}_n^{*1}, \ldots, \hat{\theta}_n^{*B}.$$

4. These B bootstrapped estimators in 3 can then be used as approximations for the bootstrap expectation, the bootstrap variance and the bootstrap quantiles:

$$\begin{aligned}
\mathbb{E}^*[\hat{\theta}_n^*] &\approx \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_n^{*i},\\
\mathrm{Var}^*(\hat{\theta}_n^*) &\approx \frac{1}{B-1}\sum_{i=1}^{B}\Big(\hat{\theta}_n^{*i} - \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_n^{*j}\Big)^2,\\
\alpha\text{-quantile of distribution of } \hat{\theta}_n^* &\approx \text{empirical } \alpha\text{-quantile of } \hat{\theta}_n^{*1}, \ldots, \hat{\theta}_n^{*B}.
\end{aligned}$$

The definition of the bootstrap values $\mathbb{E}^*$, Var^* or the bootstrap distribution are discussed next.

5.2.2 The bootstrap distribution

The bootstrap distribution, denoted here by P^*, is the conditional probability distribution which is induced by i.i.d. resampling of the data

$$Z_1^*, \ldots, Z_n^* \text{ i.i.d. } \sim \hat{P}_n, \tag{5.2}$$

given the original data. The fact that we condition on the data allows to treat the bootstrap resampling distribution $\hat{P}_n$, which is the empirical distribution of the data, as a fixed distribution.

Therefore, the bootstrap distribution P^* of $\hat{\theta}_n^* = g(Z_1^*, \ldots, Z_n^*)$ is the distribution which arises when resampling with $\hat{P}_n$ in (5.2) and applying the function g on such a bootstrap sample, exactly as done by simulation in section 5.2.1. (From a theoretical point of view, the bootstrap distribution P^* can be represented by a multinomial distribution that contains the information which and how many times of the original data appears again in the bootstrap sample (5.2)).

The bootstrap expectation of $\hat{\theta}_n^*$ is then denoted by $\mathbb{E}^*[\hat{\theta}_n^*]$ which is a conditional expectation given the data. Likewise, $\mathrm{Var}^*(\hat{\theta}_n^*)$ is a conditional variance given the data. Since $\hat{P}_n$ in (5.2) is "close" to the true data-generating probability P, the bootstrap values are "reasonable" estimates for the true quantities. For example, we can use

$$\widehat{\mathrm{Var}(\hat{\theta}_n)} = \mathrm{Var}^*(\hat{\theta}_n^*).$$

This estimate is approximately computed as in step 4 of the bootstrap algorithm in section 5.2.1. Whether such a bootstrap estimate is consistent will be discussed in the following section.

5.2.3 Bootstrap confidence interval: a first approach

The bootstrap is called to be consistent for $\hat{\theta}_n$ if, for all x,

$$\mathbb{P}[a_n(\hat{\theta}_n - \theta) \le x] \; - \; \mathbb{P}^*[a_n(\hat{\theta}_n^* - \hat{\theta}_n) \le x] \;\; \xrightarrow{P} \;\; 0 \;(n \to \infty). \tag{5.3}$$

In classical situations, $a_n = \sqrt{n}$: for example, the maximum-likelihood estimator $\hat{\theta}_n$ satisfies under regularity assumptions

$$\sqrt{n}(\hat{\theta}_{n,\mathrm{MLE}} - \theta) \xrightarrow{D} \mathcal{N}(0, I^{-1}(\theta)) \ (n \to \infty),$$

where $I(\theta)$ denotes the Fisher information at θ. Bootstrap consistency (5.3) then means

$$\sqrt{n}(\hat{\theta}^*_{n,\mathrm{MLE}} - \hat{\theta}_n) \xrightarrow{D^*} \mathcal{N}(0, I^{-1}(\theta)) \ \text{ in probability } \ (n \to \infty).$$

Consistency of the bootstrap typically holds if the limiting distribution of $\hat{\theta}_n$ is Normal, and if the data $Z_1, \ldots, Z_n$ are i.i.d.

Consistency of the bootstrap (usually) implies consistent variance and bias estimation:

$$\frac{\mathrm{Var}^*(\hat{\theta}^*_n)}{\mathrm{Var}(\hat{\theta}_n)} \xrightarrow{P} 1,$$
$$\frac{\mathbb{E}^*[\hat{\theta}^*_n] - \hat{\theta}_n}{\mathbb{E}[\hat{\theta}_n] - \theta} \xrightarrow{P} 1.$$

Moreover, consistent confidence intervals can be constructed. A two-sided confidence interval with coverage $1 - \alpha$ for a parameter θ is given by

$$\left[\hat{\theta}_n - q_{1-\alpha/2},\ \hat{\theta}_n - q_{\alpha/2}\right], \qquad \text{where } q_\alpha = \alpha\text{-quantile of } \hat{\theta}_n - \theta.$$

This is derived using elementary calculations. In analogy, the bootstrap estimated confidence interval is then defined as

$$\begin{aligned}&\left[\hat{\theta}_n - \hat{q}_{1-\alpha/2},\ \hat{\theta}_n - \hat{q}_{\alpha/2}\right],\\ &\text{where } \hat{q}_\alpha = \alpha\text{-bootstrap quantile of } \ \hat{\theta}^*_n - \hat{\theta}_n.\end{aligned} \tag{5.4}$$

Due to invariance of the quantile:

$$\hat{q}_\alpha = q^*_\alpha - \hat{\theta}_n, \qquad \text{where } q^*_\alpha = \alpha\text{-bootstrap quantile of } \ \hat{\theta}^*_n.$$

Therefore, the bootstrap confidence interval in (5.4) becomes

$$\left[2\hat{\theta}_n - q^*_{1-\alpha/2},\ 2\hat{\theta}_n - q^*_{\alpha/2}\right]. \tag{5.5}$$

Note that this is **not the same as simply taking the quantiles of the bootstrap distribution**, i.e., the simple-minded $\left[q^*_{\alpha/2},\ q^*_{1-\alpha/2}\right]$ is "backwards" and often less appropriate. The derivation which we gave for this "unintuitive" fact hinges on the consistency of the bootstrap in (5.3).

Better bootstrap confidence intervals than (5.4) exist and often have better coverage accuracy — at the price of being somewhat more difficult to implement. See also the double bootstrap below (section 5.3)

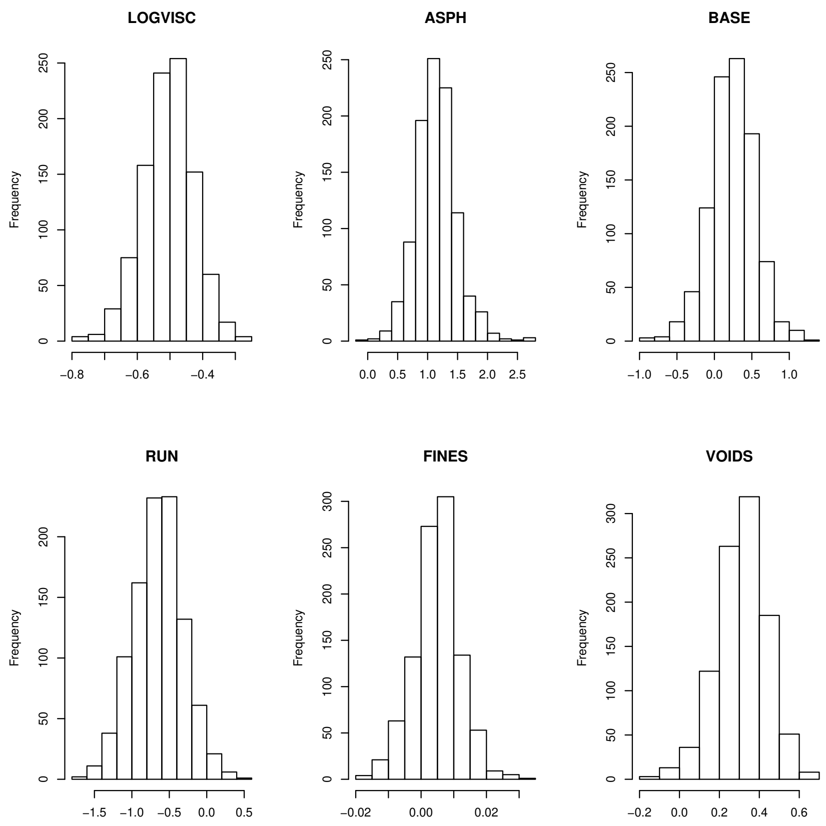

Figure 5.1: Histograms of $B = 1000$ bootstrap values $\hat{\beta}_j^*$ for the 6 coefficients in the linear model about asphalt quality from section 1.5.1.

An example

We consider here the example about asphalt quality from section 1.5.1, where a linear model has been used to model the log of rutting as a function of 6 predictor variables.

The bootstrap distributions for the 6 coefficients corresponding to the 6 predictor variables are exhibited in Figure 5.1.

The standard confidence intervals based on normally distributed data (or on asymptotic considerations) are (including an intercept)

$$\hat{\beta}_j \pm t_{n-p-1;1-\alpha/2} \cdot \widehat{s.e.}(\hat{\beta}_j),$$

where $t_{n-p-1;\alpha}$ denotes the α-quantile of a t_{n-p-1} distribution (`qt(`α`,n-p-1)`). In this example, $n = 31$, $p = 6$ when using $\alpha = 0.05$, we get $t_{n-7;1-\alpha/2} = 2.063899$. The confidence intervals are then given in Table 5.2.3.

method	intercept	LOGVISC	ASPH	BASE	RUN	FINES	VOIDS
classical	[-10.86,-0.71]	[-0.66,-0.36]	[0.60,1.70]	[-0.44,0.91]	[-1.23,-0.01]	[-0.02,0.02]	[0.09,0.54]
bootstrap	**[-11.97,0.99]**	[-0.67,-0.36]	[0.41,1.78]	[-0.37,0.87]	**[-1.26,0.09]**	[-0.01,0.02]	[0.10,0.60]

Table 5.1: Classical 95% confidence intervals and 95% bootstrap confidence intervals (using $B = 1000$) from (5.4) for the 7 coefficients in a linear model for the asphalt dataset from section 1.5.1. In bold are the bootstrap confidence intervals which suggest non-significance while the classical confidence intervals (or t-tests) would suggest significance.

5.2.4 Bootstrap estimate of the generalization error

Consider the problem of estimating the generalization error

$$\mathbb{E}[\rho(Y_{new}, \hat{m}(X_{new}))]$$

in (4.1), where ρ is a loss function such as $\rho(y,m) = |y-m|^2$ and $\hat{m}(\cdot)$ could be a regression estimator.

The bootstrap approach to estimate the generalization error is from a conceptual view as follows.

1. Generate $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*), (X_{new}^*, Y_{new}^*)$ i.i.d. $\sim \hat{P}_n$,

   where $\hat{P}_n$ is the empirical distribution of the original data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

2. Compute the bootstrapped estimator $\hat{m}^*(\cdot)$ based on $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$.

3. Compute the bootstrap generalization error

   $$\mathbb{E}^*[\rho(Y_{new}^*, \hat{m}^*(X_{new}^*))],$$

   where $\mathbb{E}^*$ is with respect to all the bootstrap variables train$^* = (X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ **and** test$^* = (X_{new}^*, Y_{new}^*)$. Use this as an estimate of the true generalization error.

The bootstrap generalization error can be re-written as follows:

$$\begin{aligned}
\mathbb{E}^*[\rho(Y_{new}^*, \hat{m}(X_{new}^*))] &= \mathbb{E}^*_{\text{train}^*}\left[\mathbb{E}^*_{\text{test}^*}[\rho(Y_{new}^*, \hat{m}(X_{new}^*)) \mid \text{train}^*]\right] \qquad (5.6)\\
&= \mathbb{E}^*_{\text{train}^*}[n^{-1}\sum_{i=1}^n \rho(Y_i, \hat{m}^*(X_i))] = n^{-1}\sum_{i=1}^n \mathbb{E}^*[\rho(Y_i, \hat{m}^*(X_i))].
\end{aligned}$$

The first equality on the second line follows because: (i) test$^* = (X_{new}^*, Y_{new}^*)$ is independent (with respect to the bootstrap distribution) from train*, and hence the inner conditional expectation is a non-conditional expectation using $\hat{m}^*(\cdot)$ as fixed (non-random because we condition on the bootstrap training data train*); (ii) the bootstrap expectation $\mathbb{E}^*[g(X^*, Y^*)] = n^{-1}\sum_{i=1}^n g(X_i, Y_i)$ for any function $g(\cdot)$.

Therefore, the bootstrap generalization error as represented in (5.6) is the average of the bootstrap errors over the observed original data (X_i, Y_i). In particular, there is no need to generate (X_{new}^*, Y_{new}^*) as conceptually introduced above in step 1. The practical algorithm then looks as follows.

1. Generate $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ by resampling with replacement from the original data.

2. Compute the bootstrapped estimator $\hat{m}^*(\cdot)$ based on $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$.

3. Evaluate err$^* = n^{-1}\sum_{i=1}^n \rho(Y_i, \hat{m}^*(X_i))$.

4. Repeat steps 1–3 B times to obtain err$^{*1}, \ldots,$ err^{*B}. Approximate the bootstrap generalization error in (5.6) by

   $$B^{-1}\sum_{i=1}^B \text{err}^{*i},$$

   and use it as an estimate for the true generalization error in (4.1).

5.3 Double bootstrap

We describe here how the bootstrap can be used twice (or multiple times) aiming to improve a bootstrap confidence interval. The same idea can also be used for potentially improving other bootstrap estimates than confidence intervals.

Suppose we wish to construct a $(1-\alpha)$-confidence interval for a parameter θ based on some estimator $\hat{\theta}$. The bootstrap interval $I^*(1-\alpha)$, as defined in (5.4), is not exact, i.e.,

$$\mathbb{P}[\theta \in I^*(1-\alpha)] = 1 - \alpha + \Delta_n,$$

with some approximation error Δ_n which will diminish as $n \to \infty$.

The main idea is now as follows: when changing the nominal coverage to $1-\alpha'$ and using $I^*(1-\alpha')$, we can get an exact actual coverage

$$\mathbb{P}[\theta \in I^*(1-\alpha')] = 1 - \alpha.$$

The problem is that α' is unknown. But another level of bootstrap can be used to **estimate** α', denoted by $\widehat{\alpha'}$, which typically achieves

$$\mathbb{P}[\theta \in I^*(1-\widehat{\alpha'})] = 1 - \alpha + \Delta_n',$$

where the new approximation error Δ_n' is typically smaller than Δ_n above.

A second level of bootstrap

In case where the original data $Z_1, \ldots, Z_n$ are replaced by n i.i.d. bootstrap realizations $Z_1^*, \ldots, Z_n^*$, we can get an exact answer for the level α' above. We can proceed in analogy to the setting above, step by step.

First, suppose the data is $Z_1^*, \ldots, Z_n^*$. By using the bootstrap for the bootstrap data $Z_1^*, \ldots, Z_n^*$, i.e., a second level bootstrap with $Z_1^{**}, \ldots, Z_n^{**}$ (see below), we can construct a confidence interval $I^{**}(1-\alpha)$ as in (5.4). We can now inspect the actual coverage for this second level bootstrap interval:

$$\mathbb{P}^*[\hat{\theta}_n \in I^{**}(1-\alpha)] \quad =: \quad h^*(1-\alpha)$$

for some function $h^* : [0,1] \to [0,1]$ which is increasing. Now use

$$1 - \alpha'^* = h^{*-1}(1-\alpha) \tag{5.7}$$

so that

$$\mathbb{P}^*[\hat{\theta}_n \in I^{**}(1-\alpha'^*)] = h^*(h^{*-1}(1-\alpha)) \quad = \quad 1-\alpha$$

is an exact confidence interval "in the bootstrap world" for the "parameter" $\hat{\theta}_n$. Therefore, the bootstrap estimate for the adjusted nominal coverage level is

$$\widehat{1-\alpha'} = 1 - \alpha'^* \text{ from (5.7).}$$

Computation of bootstrap adjusted nominal level $1 - \alpha'^*$

We can use a double (two-level) bootstrap scheme to approximately compute the value $1 - \alpha'^*$ in (5.7).

1. Draw a bootstrap sample $Z_1^*, \dots, Z_n^*$ by resampling with replacement.

   (a) Compute a bootstrap interval. That is, generate a second level bootstrap sample
   $$Z_1^{**}, \dots, Z_n^{**}$$
   by resampling with replacement from $Z_1^*, \dots, Z_n^*$. Then, analogous to (5.4), construct the double bootstrap confidence interval
   $$I^{**}(1-\alpha) = \left[\hat{\theta}^* - \hat{q}^*_{1-\alpha/2},\ \hat{\theta}^* - \hat{q}^*_{\alpha/2}\right],$$
   where $\hat{q}^*_\alpha = \alpha -$ quantile of $\hat{\theta}^{**} - \hat{\theta}^*$, $\hat{\theta}^{**} = g(Z_1^{**}, \dots, Z_n^{**})$. This is computed by repeating step 1(a) B times to approximate $\hat{q}^*_\alpha$ by the empirical α-quantile of $\hat{\theta}^{**1} - \hat{\theta}^*, \dots, \hat{\theta}^{**B} - \hat{\theta}^*$.

   (b) Evaluate whether the "parameter" $\hat{\theta}$ in the "bootstrap world" is in $I^{**}(1-\alpha)$: i.e., consider
   $$\text{cover}^*(1-\alpha) = \mathbf{1}_{[\hat{\theta} \in I^{**}(1-\alpha)]}.$$

2. Repeat M times all of step 1 to obtain $\text{cover}^{*1}(1-\alpha), \dots, \text{cover}^{*M}(1-\alpha)$. This amounts to M first level bootstrap replications $\mathbf{Z}^{*m} = Z_1^{*m}, \dots, Z_n^{*m}$ $(m = 1, \dots, M)$ and for **each** $\mathbf{Z}^{*m}$, we run B second level bootstrap replications for step 1(a). Use
   $$p^*(\alpha) := M^{-1} \sum_{i=1}^{M} \text{cover}^{*i}(1-\alpha) \tag{5.8}$$
   as an approximation for $\mathbb{P}^*[\hat{\theta} \in I^{**}(1-\alpha)]$.

3. Vary α (in all of step 1 and 2!) to find α'^* such that
   $$p^*(\alpha'^*) = 1 - \alpha \ \ (\text{the desired nominal level}) \quad \text{and use} \quad \widehat{1-\alpha'} = 1 - \alpha'^*.$$
   The search for α'^* (a "zero finding problem") can be done on a grid and/or by using a bisection strategy.

The total amount of computation requires $B \cdot M$ bootstrap samples. In case where the bootstrap interval in (5.4) is computed with B bootstrap samples, and hence also the interval I^{**} in step 1(a), the adjustment with the double bootstrap may be less important and it is then reasonable to use $M < B$ since the magnitude of M only determines the approximation for computing the actual level $\mathbb{P}^*[\hat{\theta} \in I^{**}(1-\alpha)]$ (for I^{**} computed with B bootstrap replications).

An example

We illustrate now the double bootstrap for confidence intervals in curve estimation. Figure 5.2 displays the data, having sample size $n = 100$, and a curve estimator.

Figure 5.3 then shows how the double bootstrap is used to estimate the actual coverage: displayed is an approximation of $\mathbb{P}^*[\hat{\theta}_n \in I^{**}(1-\alpha)]$ for various nominal levels $1-\alpha$. It also indicates the values for the corrected levels $1-\alpha'^*$ and it also demonstrates the effect when using a double-bootstrap corrected confidence interval instead of an ordinary interval.

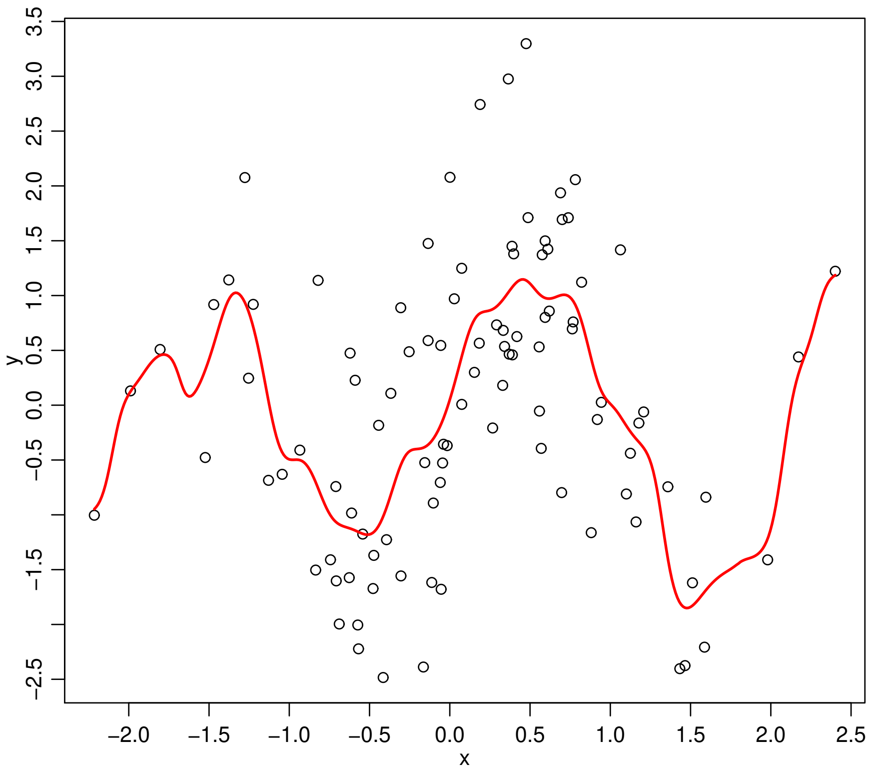

Figure 5.2: Data ($n = 100$) and estimated curve (red) using a Nadaraya Watson Gaussian kernel estimator with bandwidth $h = 0.25$.

5.4 Model-based bootstrap

Efron's nonparametric bootstrap can be viewed as simulating from the empirical distribution $\hat{P}_n$: that is, we simulate from a very general estimated nonparametric model, where the model says that the data is i.i.distributed with an unknown distribution P.

5.4.1 Parametric bootstrap

Instead of such a general nonparametric model, we sometimes assume that the data are realizations from

$$Z_1, \ldots, Z_n \text{ i.i.d. } \sim P_\theta,$$

where P_θ is given up to an unknown parameter (vector) θ.

As one among very many examples: the data could be real-valued assumed to be from the parametric model

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma^2),\ \theta = (\mu, \sigma^2).$$

In order to simulate from the parametric model, we first estimate the unknown parameter θ by $\hat{\theta}$ such as least squares in regression or maximum likelihood in general. The parametric bootstrap then proceeds by using

$$Z_1^*, \ldots, Z_n^* \text{ i.i.d. } \sim P_{\hat{\theta}},$$

instead of (5.2). Everything else, e.g. construction of confidence intervals, can then be done exactly as for Efron's nonparametric bootstrap.

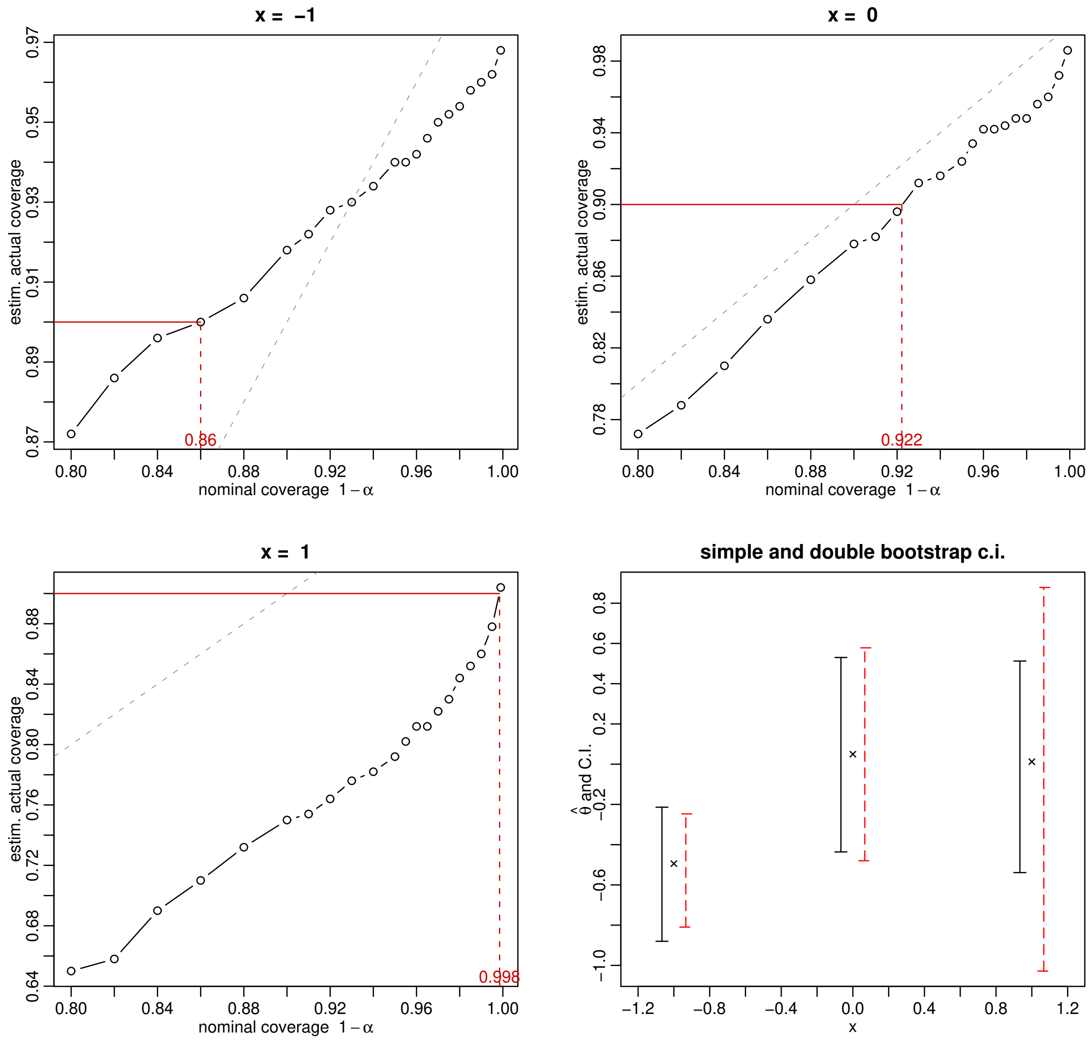

Figure 5.3: Double bootstrap confidence intervals for nonparametric curve at three predictor points $x \in \{-1, 0, 1\}$. The data ($n = 100$) and estimated curve are shown in Figure 5.2. The first three panels show the estimated actual coverages ($p^*(\alpha)$) of a bootstrap confidence interval by using the double bootstrap. The values $1 - \alpha'^*$ (for actual level $1 - \alpha = 0.9$) are 0.86, 0.922, 0.998 for the points $x = -1, 0, 1$, respectively. The fourth panel shows the ordinary bootstrap confidence intervals (solid line) and the double bootstrap corrected versions (dotted line, in red) for $x \in \{-1, 0, 1\}$. The double bootstrap was used with $B = 1000$ and $M = 500$.

Advantages and disadvantages

Why should we choose the parametric instead of the nonparametric bootstrap? The answer is "classical": if the parametric model is a very good description for the data, then the parametric bootstrap should yield more accurate variance estimates or confidence intervals since $P_{\hat{\theta}}$ is then "closer" to the true data-generating P than the nonparametric empirical distribution $\hat{P}_n$. Particularly when sample size n is small, the nonparametric estimate $\hat{P}_n$ may be poor. On the other hand, the nonparametric bootstrap is not (or less) sensitive to model-misspecification.

5.4.2 Model structures beyond i.i.d. and the parametric bootstrap

Linear model with fixed predictors

For example, a linear model with fixed predictors $x_i \in \mathbb{R}^p$ and Gaussian errors

$$\begin{aligned} &Y_i = \beta^\intercal x_i + \varepsilon_i \ (i = 1, \ldots, n), \\ &\varepsilon_1, \ldots, \varepsilon_n \ \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2),\ \theta = (\beta, \sigma^2) \end{aligned}$$

is a parametric model. The bootstrap sample can then be constructed as follows:

1. Simulate $\varepsilon_1^*, \ldots, \varepsilon_n^*$ i.i.d. $\sim \mathcal{N}(0, \hat{\sigma}^2)$.
2. Construct
$$Y_i^* = \hat{\beta}^\intercal x_i + \varepsilon_i^*,\ i = 1, \ldots, n.$$
The parametric bootstrap regression sample is then
$$(x_1, Y_1^*), \ldots, (x_n, Y_n^*),$$
where the predictors x_i are as for the original data.

Autoregressive models for time series

A Gaussian autoregressive model of order p for stationary time series is

$$\begin{aligned} &X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t \ (t = 1, \ldots, n), \\ &\varepsilon_1, \ldots, \varepsilon_n \ \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2), \end{aligned}$$

where $X_t \in \mathbb{R}$. Such a model produces correlated observations and is widely used for describing time-dependent observations. Parametric bootstrapping can then be done as follows:

1. Generate $\varepsilon_1^*, \ldots, \varepsilon_{n+m}^*$ i.i.d. $\sim \mathcal{N}(0, \hat{\sigma}^2)$ with $m \approx 1000$.
2. Construct recursively, starting with $X_0^* = X_{-1}^* = \ldots = X_{-p+1}^* = 0$,
$$X_t^* = \sum_{j=1}^{p} \hat{\phi}_j X_{t-j}^* + \varepsilon_t^*,\ t = 1, \ldots, n+m.$$
3. Use the bootstrap sample
$$X_{m+1}^*, \ldots, X_{n+m}^*.$$
The reason to throw away the first values $X_1^*, \ldots X_m^*$ is to obtain a bootstrap sample which is approximately a stationary process (by choosing m large, the arbitrary starting values in step 2 will be almost forgotten).

5.4.3 The model-based bootstrap for regression

A compromise between Efron's non- and the parametric bootstrap for regression is given by assuming possibly non-Gaussian errors. The model for the original data is

$$
\begin{aligned}
&Y_i = m(x_i) + \varepsilon_i,\\
&\varepsilon_1, \ldots, \varepsilon_n \ \text{i.i.d.} \ \sim P_\varepsilon,
\end{aligned}
$$

where P_ε is unknown with expectation 0. The regression function $m(\cdot)$ may be parametric or nonparametric. The model-based bootstrap works then as follows:

1. Estimate $\hat{m}$ from the original data and compute the residuals $r_i = Y_i - \hat{m}(x_i)$.
2. Consider the centered residuals $\tilde{r}_i = r_i - n^{-1}\sum_{i=1}^n r_i$. In case of linear regression with an intercept, the residuals are already centered. Denote the empirical distribution of the centered residuals by $\hat{P}_{\tilde{r}}$.
3. Generate
$$\varepsilon_1^*, \ldots, \varepsilon_n^* \ \text{i.i.d.} \ \sim \hat{P}_{\tilde{r}}.$$
Note that $\hat{P}_{\tilde{r}}$ is an estimate of P_ε.
4. Construct the bootstrap response variables
$$Y_i^* = \hat{m}(x_i) + \varepsilon_i^*, \ i = 1, \ldots, n,$$
and the bootstrap sample is then $(x_1, Y_1^*), \ldots, (x_n, Y_n^*)$.

Having the bootstrap sample from step 4, we can then proceed as for Efron's nonparametric bootstrap for constructing variance estimates or confidence intervals.

The advantage of the model-based bootstrap is that we do not rely on a Gaussian error assumption. The same discussion then applies about advantages and disadvantages as in section 5.4.1.