Computer-intensive methods in statistical analysis *

Dimitris N. Politis
University of California at San Diego
dpolitis@ucsd.edu

1. Resampling and the bootstrap

As far back as the late 70s, the impact of affordable, high-speed computers on the theory and practice of modern statistics was recognized by Efron [14] [15]. As a result, the bootstrap and other computer-intensive statistical methods (such as subsampling and the jackknife) have been developed extensively since that time, and now constitute very powerful (and intuitive) tools to do statistics with. The goal of this article is to provide a readable, self-contained introduction to the bootstrap and jackknife methodology for statistical inference; in particular, the focus is on the derivation of confidence intervals in general situations. In addition, a guide to the available bibliography on bootstrap methods is offered in Section 4.

1.1 The general nonparametric set-up. Suppose that $\mathbf{X} = (X_1, \ldots, X_N)$ is an independent, identically distributed (i.i.d.) sample from a population with distribution F. In other words, $F(x) = Prob(X_i \leq x)$, for $i = 1, \ldots, N$, where x is any real number; the function F is usually called a probability distribution function, or a cumulative distribution function. The sample is studied in order to estimate a certain parameter $\theta(F)$ associated with the distribution F whose form is unknown. A statistic $T = T(\mathbf{X})$ might be used to estimate $\theta(F)$ from the data. However, *a measure of the statistical accuracy of the point estimator $T(\mathbf{X})$ is also desired.* In other words, although it is an unfortunate fact of life that our estimator will not equal $\theta(F)$ exactly, the deviation of $T(\mathbf{X})$ from $\theta(F)$, i.e., the 'error' in estimating $\theta(F)$ by $T(\mathbf{X})$, could be statistically quantified; in that case, the practitioner would be able to gauge how much importance to attach to the individual 'measurement' $T(\mathbf{X})$. For example, the bias (also known as the 'systematic error') and the variance (which is responsible for the 'random error') of the estimator T are of interest, and are defined as follows:

*This paper has appeared in IEEE Signal Proc. Mag. in Jan 1998, pp 39-55.

$$Bias_F(T) = E_F T(\mathbf{X}) - \theta(F) \tag{1}$$

$$Var_F(T) = E_F T^2(\mathbf{X}) - [E_F T(\mathbf{X})]^2 \tag{2}$$

where E_F denotes expectation under the F distribution. The quantity $T(\mathbf{X}) - \theta(F)$, i.e., the estimator minus the estimand, represents the 'error' in estimating $\theta(F)$ by $T(\mathbf{X})$; it is also sometimes called a 'root' [13], [5]. Much (if not most) of statistical theory and practice is devoted to studying the sampling properties of such 'roots'; in particular, bootstrap methods provide easy-to-use, and rather powerful tools for this purpose.

To fix ideas, consider for example the case where $\theta(F)$ is a location parameter, say the mean or median of F, and $T(\mathbf{X})$ is the corresponding sample statistic (sample mean, sample median, etc.); nonetheless, our discussion is general, and not at all limited to the simple location problem. In many practical situations the Central Limit Theorem can be invoked to assert that the estimator $T(\mathbf{X})$ is approximately distributed as a Gaussian random variable. This will typically be true for most 'good' estimators, provided the sample size N is large enough, in which case the estimator is said to be *asymptotically* normal, and an approximate interval estimate, i.e., a *confidence interval*, for $\theta(F)$ can be formed, in addition to the point estimate $T(\mathbf{X})$.

1.2 Confidence intervals based on asymptotic normality. If the bias $Bias_F(T)$ is negligible (compared to the square root of the variance $Var_F(T)$), a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ will be of the usual form

$$[T(\mathbf{X}) - z\sqrt{Var_F(T)}, T(\mathbf{X}) + z\sqrt{Var_F(T)}], \tag{3}$$

where $z = z(1-\alpha/2)$ is the $1-\alpha/2$ quantile[1] of the standard normal distribution. If $Bias_F(T)$ is not negligible, the confidence interval must be adjusted appropriately; generally, a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ will be given by

$$[T(\mathbf{X}) - Bias_F(T) - z\sqrt{Var_F(T)}, T(\mathbf{X}) - Bias_F(T) + z\sqrt{Var_F(T)}] \tag{4}$$

Note that the aforementioned confidence intervals are based on the fact that the shape of the large-sample distribution of the root $T(\mathbf{X})-\theta(F)$ is known; it is the bell-shaped normal. However, to formulate this confidence interval one needs to know $Bias_F(T)$ and $Var_F(T)$.

[1] All probability (cumulative) distribution functions are monotone increasing. If a distribution $F(x)$ is *strictly* increasing, i.e., $x < y$ implies $F(x) < F(y)$, then its α *quantile* is given by $F^{-1}(\alpha)$, where F^{-1} is the inverse function of F; for example, the normal distribution is strictly monotone. If F happens *not* to be strictly increasing, then it must have some regions where its graph is flat; in that case, the α quantile of F is defined as the smallest x-value such that $F(x) \geq \alpha$.

Estimates of $Bias_F(T)$ and $Var_F(T)$ might be available in the statistical literature for different problems. For example, if $T(\mathbf{X}) = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the sample mean, and $\theta(F) = E_F X_1$ is the population mean, then it is well known that $Bias_F(T) = 0$, and $Var_F(T) = \frac{1}{N} Var_F(X_1)$, where $Var_F(X_1)$ can be estimated by the sample variance $\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$. If $T(\mathbf{X})$ is the sample median and $\theta(F)$ is the population median, estimates of $Bias_F(T)$ and $Var_F(T)$ can still be calculated (cf. [32] p. 354)), but are substantially more complicated. As a matter of fact, to estimate the large-sample variance of the sample median, one needs to estimate the value of the derivative of F (i.e., the probability *density* function F' –assuming that it exists) *at the location of the true median*; note that nonparametric density estimation is a difficult issue, and it would be nice if it could be somehow by-passed, especially in such a 'bread-and-butter' everyday example as the sample median. The bootstrap (and the closely related jackknife [14], [16]) come to our rescue here by providing alternative methods to *easily* obtain estimates of $Bias_F(T)$ and $Var_F(T)$ for a wide variety of statistics $T(\mathbf{X})$. However, before going into that, let us look at this problem from a different angle.

1.3 The usefulness of Monte Carlo randomization. Let us suppose, for the sake of argument, that the population and its distribution F are in fact known – a very unrealistic assumption in practice. In that case, $Bias_F(T)$ and $Var_F(T)$ could be calculated exactly by analytical methods, or approximately by Monte Carlo simulation, in case the analytical computation is difficult.

The idea behind Monte Carlo simulation is the following. Since the population is considered known, we can draw any number of i.i.d. samples from it. Suppose that we draw B samples, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(B)}$, where each sample consists of N i.i.d. observations from the population F. If B is large enough, the strong law of large numbers can be invoked to claim that

$$E_F g(T(\mathbf{X})) \simeq \frac{1}{B}\sum_{i=1}^{B} g(T(\mathbf{X}^{(i)})) \tag{5}$$

where $g(\cdot)$ is some function, e.g. $g(x) = x$ or $g(x) = x^2$. Then we would have

$$Bias_F(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{(i)}) - \theta(F) \tag{6}$$

$$Var_F(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T^2(\mathbf{X}^{(i)}) - [\frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{(i)})]^2. \tag{7}$$

In other words, since bias and variance give us a rough idea about how the values of the statistic T vary (fluctuate) *across samples*, we estimate $Bias_F(T)$ and $Var_F(T)$ by the empirically observed (over our artificially generated samples, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(B)}$) bias and variance.

Therefore, the idea is that we can estimate (by Monte Carlo) the variability of the statistic T across samples by looking at the empirically observed variability of T across *our* (artificially generated) samples. Thus, we could also estimate (by Monte Carlo) the whole sampling distribution of the root $T(\mathbf{X}) - \theta(F)$ without reference to the asymptotic (for large N) normal distribution.

Define $P_F(A)$ to be the probability of event A occurring, under the assumption that the population has distribution F, and let

$$Dist_{T-\theta,F}(x) \equiv P_F(T(\mathbf{X}) - \theta(F) \leq x). \tag{8}$$

Again, although F is considered known, the analytical evaluation of $Dist_{T-\theta,F}(x)$ may be difficult, and we may resort to Monte Carlo. Observe that $Dist_{T-\theta,F}(x)$ is just a shifted (centered) version of

$$Dist_{T,F}(x) = P_F(T(\mathbf{X}) \leq x) \tag{9}$$

so that $Dist_{T-\theta,F}(x) = Dist_{T,F}(x + \theta(F))$. If we define the indicator function of event A by the formula

$$\mathbf{1}(A) = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{else} \end{cases}$$

then, using equation (5) with $g(T(\mathbf{X})) = \mathbf{1}(T(\mathbf{X}) \leq x)$, and the fact that $E_F\mathbf{1}(A) = P_F(A)$, we have [2]

$$Dist_{T,F}(x) \simeq \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}(T(\mathbf{X}^{(i)}) \leq x) = \frac{1}{B}(\#T(\mathbf{X}^{(i)}) \leq x) \tag{10}$$

i.e., the theoretical probability should be approximately equal to the observed sample proportion if B is large.

Knowledge of $Dist_{T-\theta,F}(x)$, for all real x, would immediately yield a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ in the form

$$[T(\mathbf{X}) - q(1-\alpha/2), T(\mathbf{X}) - q(\alpha/2)], \tag{11}$$

where $q(\alpha/2)$ and $q(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{T-\theta,F}(x)$ distribution respectively. The above confidence interval is equal-tailed, meaning that the probability the interval's left end-point is bigger than $\theta(F)$ is equal to the probability the interval's right end-point is smaller than $\theta(F)$. Other constructions (e.g., symmetric, shortest length, etc.) for confidence intervals are also available (cf. [24], [25]) and possess some interesting theoretical properties; nevertheless, the confidence intervals that are most often used in practice are equal-tailed [20].

[2] Note that $(\#T(\mathbf{X}^{(i)}) \leq x)$ reads: number of the $T(\mathbf{X}^{(i)})$'s among $T(\mathbf{X}^{(1)}), \ldots, T(\mathbf{X}^{(B)})$ that are observed to be less or equal to x; equation (10) should be viewed as describing a function of the real argument x, and can be plotted as such.

Now from equation (10), the quantiles of $Dist_{T,F}(x)$ (and therefore also those of $Dist_{T-\theta,F}(x)$)) can be approximately calculated, and the confidence interval (11) constructed with the help of our Monte Carlo Simulation.

1.4 The bootstrap principle. To summarize, *if* the population and its distribution F *were* known, then we would be able to calculate (analytically or by Monte Carlo simulations) $Bias_F(T)$, $Var_F(T)$, and $Dist_{T-\theta,F}(x)$. However, in the practical problem the population and its distribution F are *not* known. The bootstrap method now is an outcome of the following simple idea: *since you do not have the whole population, do the best with what you* do *have, which is the observed sample* $\mathbf{X} = (X_1, \ldots, X_N)$.

In other words, the bootstrap method amounts to treating your observed sample as *if* it *exactly* represented the whole population; see the pioneering paper by Efron [14]. In this fashion, the Monte Carlo procedure in which B i.i.d. samples were drawn from the population is modified to read:

- Draw B i.i.d. samples $\mathbf{X}^{*(1)}, \ldots, \mathbf{X}^{*(B)}$ (each of size N) from the sample population consisting of the observations $\{X_1, \ldots, X_N\}$. In the bootstrap terminology, these B i.i.d. samples are called *resamples*. Of course, drawing an i.i.d. sample from a finite population such as $\{X_1, \ldots, X_N\}$, amounts to sampling with replacement from the set $\{X_1, \ldots, X_N\}$.

Note that, as the whole population has distribution F, the sample population has distribution $\hat{F}$, the so-called *empirical* distribution, which is defined as

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(X_i \leq x) = \frac{1}{N}(\#X_i \leq x), \tag{12}$$

for any real number x. To elaborate, in order to form the ith resample $\mathbf{X}^{*(i)} = (X_1^{*(i)}, \ldots, X_N^{*(i)})$, we sample with replacement from the set $\{X_1, \ldots, X_N\}$, or, using a different terminology, we *take an i.i.d. sample of size N from a population with distribution* $\hat{F}$.

1.5 The bootstrap as a 'plug-in' method. This last observation suggests a different perspective for the implementation of the bootstrap as a simple '*plug-in*' method. Namely, if at a certain formula the unknown distribution F appears, you just substitute $\hat{F}$ in place of F to get its bootstrap approximation. For example, the bootstrap approximations to $Bias_F(T)$ and $Var_F(T)$ are given simply by

$$Bias^*(T) = Bias_{\hat{F}}(T) \tag{13}$$

$$Var^*(T) = Var_{\hat{F}}(T). \tag{14}$$

It should be noted that $\theta(\hat{F})$ is just the sample version of the population parameter $\theta(F)$. In most cases, the statistic $T(\mathbf{X})$ is chosen to be [3] just $\theta(\hat{F})$. For example, if $\theta(F)$ is the population median, then we might want to use the sample median to estimate it, i.e. $T(\mathbf{X}) = \theta(\hat{F})$. Unless otherwise stated, we will henceforth assume that $\theta(\hat{F}) \equiv T(\mathbf{X})$ for simplicity; in a different situation the 'plug-in' principle can be appropriately modified.

As stated in the previous Subsection, $\hat{F}$ is nothing more than the distribution of the sample population. Nonetheless, $\hat{F}$ can also be viewed as our (nonparametric) estimate of the distribution F of the whole population, i.e., we can view $\hat{F}(x)$ (for some fixed real number x) as our estimate of $F(x)$; thus, the 'plug-in' bootstrap principle is nothing more than the familiar 'plug-in-an-estimator-in-place-of-an-unknown' principle used routinely throughout science and engineering!

To elaborate on treating $\hat{F}$ as an estimator of F, let us fix a real number x; if we let $U_i = \mathbf{1}(X_i \leq x)$, it is obvious that $U_1, \ldots, U_N$ are i.i.d. $Bernoulli(p)$ random variables, i.e. 1-0 coin flips, or success/failure dichotomies. Note that here $p = Prob(X_i \leq x) = F(x)$; thus, estimating this p probability (=probability of "success" of the $Bernoulli$ random variables), i.e., estimating $F(x)$, boils down to finding the observed proportion of successes among the N trials comprising our sample. The RHS of equation (12) is nothing more than this observed (sample) proportion.

1.6 A parametric set-up. The 'plug-in-an-estimator' viewpoint permits us to see how the bootstrap would work in a parametric problem as well. A parametric framework starts by postulating that the distribution $F(x)$ is known up to some parameter θ; in other words, it is postulated that F belongs to a known class of functions $\{F_\theta(\cdot) : \theta \in \Theta\}$ parametrized by θ, where Θ is the assumed parameter space.

Hence, to pin-point F we just need to know its corresponding θ-value. Equivalently, to estimate $F(x)$ from sample data, we just need to estimate its corresponding θ-value. Thus, our parametric estimator of $F(x)$ is simply $F_{\hat{\theta}}(x)$, where $\hat{\theta} = T(\mathbf{X})$ is the estimated (from our sample) value of the parameter θ.

Consequently, the parametric bootstrap method would be to approximate quantities such as $Bias_F(T)$, $Var_F(T)$, and $Dist_{T,F}(x)$ by $Bias_{F_{\hat{\theta}}}(T)$, $Var_{F_{\hat{\theta}}}(T)$, and $Dist_{T,F_{\hat{\theta}}}(x)$ respectively. All the Monte Carlo approximations remain valid, except that in the parametric set-up, to form

[3] It is interesting to note that if $\theta(F) = E_F g(X_i)$ for some function $g(\cdot)$, then θ is a *linear* function of F; i.e., if F_1, F_2 are two distributions, then $\theta(\lambda F_1 + (1-\lambda)F_2) = \lambda\theta(F_1) + (1-\lambda)\theta(F_2)$, for all $\lambda \in [0,1]$. In that case, the statistic $\theta(\hat{F})$ is called a *linear* statistic. The prime example of a linear function $\theta(.)$ and a linear statistic $\theta(\hat{F})$ is of course provided by the population mean and sample mean respectively, where $g(\cdot)$ is the identity function, i.e. $g(x) = x$.

the ith resample $\mathbf{X}^{*(i)} = (X_1^{*(i)}, \ldots, X_N^{*(i)})$, we take an i.i.d. sample from a population with distribution $F_{\hat{\theta}}$.

Note that in parametric problems, the theory of Maximum Likelihood estimation and Fisher information are traditionally used to get point and interval estimates of the unknown parameter θ (cf. [37]); however, the bootstrap will tend to give more accurate interval estimates –as compared to the standard intervals based on asymptotic normality of the Maximum Likelihood Estimators (cf. [25], [20]). Having said that, let us return and focus our attention on the general nonparametric problem, that is, the problem where F is completely unknown, since here the bootstrap is more urgently needed.

1.7 Construction of bootstrap confidence intervals. As was mentioned before, to calculate $Bias_F(T)$ and $Var_F(T)$ we might have to resort to Monte Carlo simulation even if the distribution F were known; see equations (6), (7). Thus to calculate $Bias_{\hat{F}}(T)$ and $Var_{\hat{F}}(T)$ we might use the following Monte Carlo approximations:

$$Bias^*_F(T) = Bias_{\hat{F}}(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{*(i)}) - \theta(\hat{F}) \tag{15}$$

$$Var^*_F(T) = Var_{\hat{F}}(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T^2(\mathbf{X}^{*(i)}) - [\frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{*(i)})]^2 \tag{16}$$

The above mentioned bootstrap approximations to $Bias_F(T)$ and $Var_F(T)$ can be used to yield a confidence interval for $\theta(F)$ based on the normal approximation of equation (4). Alternatively, we can by-pass the normal approximation and set confidence intervals for $\theta(F)$ based on the exact distribution of the root $T(\mathbf{X}) - \theta(F)$ given in equation (8).

Of course, this exact distribution is not known, but a bootstrap approximation is available. More specifically, the bootstrap approximation to $Dist_{T-\theta,F}(x)$ is simply given by

$$Dist^*_{T-\theta,F}(x) = Dist_{T-\theta,\hat{F}}(x) \tag{17}$$

and consequently an equal-tailed $(1-\alpha)100\%$ *bootstrap* confidence interval for $\theta(F)$ would be

$$[T(\mathbf{X}) - q^*(1-\alpha/2), T(\mathbf{X}) - q^*(\alpha/2)], \tag{18}$$

where $q^*(\alpha/2)$ and $q^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{T-\theta,\hat{F}}(x)$ distribution respectively.

It should be mentioned at this point that this just one of many possible ways to construct a bootstrap confidence interval, namely the 'percentile' method, the 'percentile-t' or 'bootstrap-t', the 'BC_a' method, etc.; see [20] (chapter 22) for a thorough discussion, and [13], [24], [25] for a

comparison of bootstrap confidence intervals. Note that in the terminology of [24], equation (18) represents a confidence interval based on the 'hybrid' method, whereas in [25], equation (18) is the 'other percentile method' confidence interval. To avoid the confusion we will refer to equation (18) simply as the *root* method for bootstrap confidence intervals.

Now we can easily evaluate the bootstrap distributions $Dist_{T-\theta,\hat{F}}(x)$ and $Dist_{T,\hat{F}}(x)$ –and therefore their quantiles as well– using a Monte Carlo simulation. In order to start in this direction, observe that we can write $Dist_{T,\hat{F}}(x) = P_{\hat{F}}(T(\mathbf{X}^*) \leq x)$, and $Dist_{T-\theta,\hat{F}}(x) = P_{\hat{F}}(T(\mathbf{X}^*) - \theta(\hat{F}) \leq x)$, where, as before, $\mathbf{X}^*$ is an i.i.d. *resample* from a population with distribution $\hat{F}$. Thus our Monte Carlo approximations to the bootstrap distributions are immediate:

$$Dist^*_{T,F}(x) = Dist_{T,\hat{F}}(x) \simeq \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}(T(\mathbf{X}^{*(i)}) \leq x) = \frac{1}{B}(\#T(\mathbf{X}^{*(i)}) \leq x) \tag{19}$$

and

$$Dist^*_{T-\theta,F}(x) = Dist_{T-\theta,\hat{F}}(x) = Dist_{T,\hat{F}}(x+\theta(\hat{F})) \simeq \frac{1}{B}(\#T(\mathbf{X}^{*(i)}) \leq x+\theta(\hat{F})). \tag{20}$$

Similarly to equation (10), the functions $Dist_{T-\theta,\hat{F}}(x)$ and $Dist_{T,\hat{F}}(x)$ should be viewed as functions of the real argument x, and can be plotted as such. Alternatively, a practitioner can plot a *histogram* of the $T(\mathbf{X}^{*(i)})$, for $i = 1, ..., B$; this histogram would be an approximation to the probability *density* function of the random variable $T(\mathbf{X})$, while $Dist_{T,\hat{F}}(x) = P_{\hat{F}}(T(\mathbf{X}^*) \leq x)$ is an approximation to the *cumulative* probability distribution function of $T(\mathbf{X})$. The article [58] contains many examples of such plotted histograms that are very helpful in terms of visual inspection and interpretation of the variability of $T(\mathbf{X})$.

Note that the approximation to $Dist_{T,\hat{F}}(x)$ as described by the RHS of equation (19) actually represents the 'empirical' distribution function of the observed $T(\mathbf{X}^{*(i)})$, or, in other words, the distribution function of the (pseudo)sample consisting of $T(\mathbf{X}^{*(i)})$, $i = 1, ..., B$. The graph of the approximation given by the RHS of (19) is of a very simple form: it looks like a 'ladder', i.e., the graph is flat, except for jumps (='steps') of size $1/B$ occurring at the locations of the observed $T(\mathbf{X}^{*(i)})$. These observations point to a very easy way of obtaining approximations to the quantiles of distribution $Dist_{T,\hat{F}}(x)$ from which the quantiles of $Dist_{T-\theta,\hat{F}}(x)$ can be readily figured out; for, if $q^*(\alpha/2)$ and $q^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{T-\theta,\hat{F}}(x)$ distribution, then

$$q^*(\alpha/2) = Q^*(\alpha/2) - \theta(\hat{F}), \qquad q^*(1-\alpha/2) = Q^*(1-\alpha/2) - \theta(\hat{F}), \tag{21}$$

where $Q^*(\alpha/2)$ and $Q^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{T,\hat{F}}(x)$ distribution.

We now describe how to easily find approximations to the $Q^*(\alpha/2)$ and $Q^*(1-\alpha/2)$ quantiles based on the RHS of (19). Start by sorting the values $T(\mathbf{X}^{*(i)})$, $i = 1, ..., B$, of the (pseudo)

sample and recording them in ascending order, i.e., $T_1^* \leq T_2^* \leq \cdots \leq T_B^*$. Now observe that, if k is a positive integer, then

$$Dist_{T,\hat{F}}(T_k^*) \simeq \frac{1}{B}(\#T(\mathbf{X}^{*(i)}) \leq T_k^*) = k/B; \tag{22}$$

this is because (by construction) exactly[4] k out of the B values of $T(\mathbf{X}^{*(i)})$ are less than (or equal to) T_k^*. Thus, the two approximate quantiles are:

$$Q^*(\alpha/2) \simeq T_{k_1}^*, \qquad Q^*(1-\alpha/2) \simeq T_{k_2}^*, \tag{23}$$

where $k_1 = \lfloor B\alpha/2 \rfloor + 1$, $k_2 = \lfloor B(1-\alpha/2) \rfloor + 1$, and $\lfloor . \rfloor$ denotes the integer part. Putting equations (21) and (23) together, we obtain a quick-and-easy alternative formulation of the root confidence interval of equation (18) as

$$[2T(X) - T_{k_2}^*, 2T(X) - T_{k_1}^*]. \tag{24}$$

Having found the $Q^*(\alpha/2)$ and $Q^*(1-\alpha/2)$ quantiles (at least approximately), we are now in the position to also formulate the equal-tailed $(1-\alpha)100\%$ *percentile* bootstrap confidence interval for $\theta(F)$; this is given simply by the interval

$$[T_{k_1}^*, T_{k_2}^*]. \tag{25}$$

The percentile bootstrap confidence intervals are also very popular (as popular as the root intervals, if not more); see, for instance, Example 1 and Table 2 in [58] where the percentile intervals are employed. However, the justification for their use is not obvious from what has been discussed so far here. In addition, comparing the root interval (24) to the percentile interval (25), we note that the roles of $T_{k_1}^*, T_{k_2}^*$ get somehow 'interchanged'.

Our Subsection 1.8 contains some discussion on how it is at least plausible that both the root and the percentile interval are valid; nevertheless, the bottom line is (see [20] p. 174)) that both the root interval (24) and the percentile interval (25) could be improved: an improvement of the root interval is given by the 'bootstrap-t' interval presented in Subsection 1.8, while an improvement of the percentile interval is given by the 'bias-corrected accelerated' BC_a interval. Rather than going into more detail regarding the percentile and BC_a intervals here, we can refer the reader to the very interesting exposition in [20], p. 170 and on).

It should also be noted that since the resampling procedure implicit in equation (20) is done with the sample $X_1, \ldots, X_N$ being *fixed* and playing the role of a population with distribution

[4]Consistently with our definition of quantile, if it so happens that $T_k^* = T_{k+1}^*$ for some k in equation (22), then this simply means that the k/B and the $(k+1)/B$ quantiles of the distribution given by the RHS of (19) happen to be the same (and both equal to $T_k^* = T_{k+1}^*$).

$\hat{F}$, the sample statistic $\theta(\hat{F})$ is just a fixed number, calculated once and for all from the original sample $X_1, \ldots, X_N$. In the bootstrap literature, the terminology is that the resampling is done *conditionally* on the data $X_1, \ldots, X_N$.

Finally, recall that the construction of confidence intervals and hypothesis testing are dual problems in statistical theory, i.e., one can perform hypothesis tests on the basis of confidence intervals and vice versa. So it should be of no surprise that the bootstrap can be used for the purposes of hypothesis testing; see [58] for more details.

1.8 Higher order accuracy of the bootstrap and studentization. The reason for the success and popularity of the bootstrap methodology is twofold: (a) it provides answers (confidence intervals, standard error estimates, etc.) in complicated situations in a straightforward, 'automatic' way, and (b) it provides *more accurate* answers in standard settings, more accurate as compared to the ubiquitous normal approximation. So far we have discussed only part (a) above; we will now focus on (b).

Suppose that we have at our disposal a consistent[5] estimator of the variance $Var_F(T)$; let us call this estimator $\widehat{Var}_F(T)$. To fix ideas, note that if the statistic $T(\mathbf{X})$ of interest is the sample mean $\bar{X}$, then there is available a simple consistent estimator of $Var_F(T)$, namely $\widehat{Var}_F(T) = s^2/N$, where $s^2 = (N-1)^{-1} \sum_{k=1}^{N} (X_k - \bar{X})^2$ is the sample variance. Dividing the statistic $T(\mathbf{X})$ by its estimated standard deviation $\sqrt{\widehat{Var}_F(T)}$ is usually referred to as 'studentization', since –*if* the data are Gaussian and *if* $T(\mathbf{X})$ happens to be the sample mean– this would result in Student's t-distribution. For a general statistic $T(\mathbf{X})$ we can also consider the sampling distribution of the 'studentized' root S_θ, i.e.,

$$Dist_{S_\theta,F}(x) = P_F\left(\frac{T(\mathbf{X}) - \theta(F)}{\sqrt{\widehat{Var}_F(T)}} \leq x\right), \tag{26}$$

where $S_\theta \equiv (T(\mathbf{X}) - \theta(F))/\sqrt{\widehat{Var}_F(T)}$. Knowledge of $Dist_{S_\theta,F}(x)$ for all real x would yield a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ in the form

$$[T(\mathbf{X}) - u(1-\alpha/2)\sqrt{\widehat{Var}_F(T)}, T(\mathbf{X}) - u(\alpha/2)\sqrt{\widehat{Var}_F(T)}], \tag{27}$$

where $u(\alpha/2)$ and $u(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{S_\theta,F}(x)$ distribution respectively.

Note however that for general statistics, or even for the sample mean if we are not willing to assume that data are Gaussian, the distribution $Dist_{S_\theta,F}(x)$ and its quantiles are unknown;

[5] An estimator is said to be *consistent* if it is large-sample accurate with high probability, i.e., if it converges (in probability) to its target value as the sample size increases.

nevertheless, $Dist_{S_\theta,F}(x)$ and its quantiles can be estimated by the bootstrap, similarly to what was discussed in the previous Subsections. In particular, the bootstrap distribution $Dist^*_{S_\theta,F}(x)$ that can be used to approximate $Dist_{S_\theta,F}(x)$ is given by

$$Dist^*_{S_\theta,F}(x) = Dist_{S_\theta,\hat{F}}(x) \simeq \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}(\#T(\mathbf{X}^{*(i)}) \leq x\sqrt{\widehat{Var}_F{}^{*(i)}(T)} + \theta(\hat{F}))$$

$$= \frac{1}{B}(\#T(\mathbf{X}^{*(i)}) \leq x\sqrt{\widehat{Var}_F{}^{*(i)}(T)} + \theta(\hat{F})), \tag{28}$$

where $\widehat{Var}_F{}^{*(i)}(T)$ is the estimate of the variance of the statistic $T(\mathbf{X})$ *as computed from the* $\mathbf{X}^{*(i)}$ *resample*. For example, in the sample mean case, $\widehat{Var}_F{}^{*(i)}(T) = (N-1)^{-1}\sum_{k=1}^{N}(X_k^{*(i)} - \bar{X}^{*(i)})^2$, where $\bar{X}^{*(i)} = N^{-1}\sum_{k=1}^{N} X_k^{*(i)}$.

Note that, if a variance estimate is *not* readily available, $\widehat{Var}_F(T)$ *itself* could be a bootstrap estimate constructed as in Subsection 1.3; in that case, $\widehat{Var}_F{}^{*(i)}(T)$ is the bootstrap variance estimate computed from the $\mathbf{X}^{*(i)}$ *resample*! In other words, we have an *iterated* or *nested* bootstrap –a bootstrap simulation for each of the original bootstrap resamples; cf. [25] or [20] for more details.

In any case, an equal-tailed $(1-\alpha)100\%$ bootstrap confidence interval for $\theta(F)$

$$[T(\mathbf{X}) - u^*(1-\alpha/2)\sqrt{\widehat{Var}_F(T)}, T(\mathbf{X}) - u^*(\alpha/2)\sqrt{\widehat{Var}_F(T)}], \tag{29}$$

where $u^*(\alpha/2)$ and $u^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap $Dist^*_{S_\theta,F}(x)$ distribution respectively; the confidence interval of equation (29) is called a bootstrap-t or a percentile-t interval due to the 'studentization'.

As it turns out [50], the confidence interval of equation (29) is *more* accurate [6] than *either* the root bootstrap interval of equation (18), *or* the normal confidence interval of equation (3); this is what is meant by 'higher order accuracy of the bootstrap'. The mathematical explanation of this phenomenon is based on the theory of Edgeworth expansions of $Dist_{S_\theta,F}(x)$ and $Dist^*_{S_\theta,F}(x)$ in powers of $1/\sqrt{N}$, and can be found in [25] and [51]. Note that this higher order accuracy comes at a price, since the iterated bootstrap is much more computer intensive than the simple bootstrap; however, in the sample mean case the extra computational burden is minuscule, because a variance estimate can be computed without Monte Carlo simulation.

Finally, let us note that the comparison of the (studentized) bootstrap to the normal approximation is not accidental. As a matter of fact, the question may be posed: "When does the bootstrap work?", i.e., under what conditions do the bootstrap confidence intervals (18), (29),

[6] Comparing different confidence intervals of *approximate* coverage level $(1-\alpha)100\%$ for a parameter θ, the confidence interval whose coverage level is closer to the nominal level $(1-\alpha)100\%$ is said to be more *accurate*; see e.g. [20], chapter 22.2).

etc. have coverage probability approximately equal to $(1-\alpha)100\%$ as they are supposed to. Although we do not attempt to give a definitive answer to this difficult question, it is important to mention that the validity of the bootstrap seems to be somehow tied to the concurrent availability of the normal approximation. [7]

Consequently, if both (the normal and the bootstrap) approximations are simultaneously valid, it is natural to ask: "which is better?", in which case the typical answer is that the "bootstrap is better than the normal" –"better" in the sense of giving more accurate confidence intervals– which explains the recent popularity of the bootstrap. Another way of understanding the "better-than-the-normal" property of the bootstrap is to think of " the normal as the "ultimate" (i.e. final) asymptotic approximation, while the bootstrap may be thought as a "penultimate" asymptotic approximation, the distinction being that the sample size required for the "ultimate" approximation to be valid is generally larger than the sample size required for the "penultimate" to be valid. In others words, the bootstrap approximation 'kicks in' before (in terms of sample size) the normal approximation; the latter kicks in "ultimately" (i.e., for very large N). Many examples can be found in the literature [20] where the bootstrap is shown to work with N being as small as 10 or 15.

To elaborate even more on the question of higher order efficiency of the bootstrap, note that in regular cases (e.g. if $T(\mathbf{X}) = \bar{X}$ is the sample mean) it can be calculated that the skewness[8] of the (true) distribution of $T(\mathbf{X})$ is approximately equal to $c_F/\sqrt{N}$, where c_F is a parameter depending on F. While the skewness of the normal approximation is zero due to its symmetry, the skewness of the bootstrap approximation is $\hat{c}_F/\sqrt{N}$, with $\hat{c}_F$ being a consistent estimator of c_F. In other words, whereas the normal approximation matches the first two moments (mean and variance) of $T(\mathbf{X})$ but misses the third moment (the skewness), the bootstrap approximation matches all three moments (mean, variance, and skewness); this is what is commonly referred to as the bootstrap 'capturing the skewness' of the limit distribution, and can be viewed intuitively

[7] If $T(\mathbf{X})$ is a linear statistic, a very interesting result of Giné and Zinn [22] shows that the bootstrap would work only if $T(\mathbf{X})$ is asymptotically normal with $Var_F(T)$ being asymptotically proportional to $1/N$; more on this property of the variance $Var_F(T)$ may be found in our Subsection 2.1. This point of view allows for a heuristic explanation regarding how the root confidence interval of equation (24), and the percentile interval of equation (25) could be (under some conditions) simultaneously valid. The reason is that, if the root $T(\mathbf{X})-\theta(\hat{F})$ is approximately normal $N(0, Var_F(T))$, then $\theta(\hat{F})-T(\mathbf{X})$ is also approximately normal $N(0, Var_F(T))$; this is due to the symmetry of the normal probability density. In other words, the large-sample distributions of $T(\mathbf{X})-\theta(\hat{F})$ and of $\theta(\hat{F})-T(\mathbf{X})$ are the *same*; thus the confidence intervals (24) and (25) are both valid (at least to first order).

[8] The *skewness* of the random variable T is defined as the standardized third central moment, i.e., skewness of T equals $E_F(T-E_FT)^3/\left(E_F(T-E_FT)^2\right)^{3/2}$. The skewness is a rough measure of the asymmetry of the distribution of T about its mean; for example, zero skewness indicates a symmetric distribution, positive skewness indicates the existence of a long right 'tail' of the distribution, and so on.

as the reason behind the "better-than-the-normal" property of the bootstrap. The fact that *both* approximations (the normal and the bootstrap) are *concurrently* valid if N is very large (i.e., "ultimately") can be explained by noting that the (true) skewness $c_F/\sqrt{N}$ vanishes for very large N; thus the normal approximation's prescription of zero as the skewness of T is also a valid approximation if N is very large.

1.9 Transformations and variance stabilization. The reader should also refer to the textbook by Efron and Tibshirani [20] for a different construction of higher order accurate bootstrap confidence intervals, the BC_a intervals, that are based on the idea of 'bias correction'. As the bootstrap-t confidence intervals can be considered a refinement and improvement over the root intervals, similarly the BC_a intervals can be thought of as a refinement and improvement over the percentile intervals. It is quite interesting to note that the BC_a intervals have the additional desirable property of being 'transformation invariant', a property shared by the percentile intervals of (25), but *not* shared by either the root intervals of (18), or the bootstrap-t intervals of (29), nor by the normal approximation interval of equation (3).

To explain the property of 'transformation invariance', consider a (strictly) monotone function $g(\cdot)$, and its inverse $g^{-1}(\cdot)$. Since $T = T(\mathbf{X})$ is considered to be a good estimator of $\theta = \theta(F)$, then it follows that $g(T)$ is a good estimator of $g(\theta)$. Suppose $[l, u]$ is an equal-tailed $(1-\alpha)100\%$ approximate confidence interval for $\theta(F)$ constructed using any of the available methods, i.e., normal theory of equation (3), root bootstrap of equation (18), percentile bootstrap of equation (25), bootstrap-t of equation (29), or bootstrap BC_a.

Observe that $g(T)$ is just a statistic based on our sample, and it can be 'bootstrapped' as well. In other words, the sampling distribution of $g(T)$ can be estimated, and an equal-tailed $(1-\alpha)100\%$ confidence interval for $g(\theta)$ can be formed, by the same method used to obtain the interval for $\theta(F)$; say this interval is $[g_l, g_u]$. It then follows that $[g^{-1}(g_l), g^{-1}(g_u)]$ is an approximate $(1-\alpha)100\%$ confidence interval for $\theta(F)$, and this new confidence interval should be compared to the interval $[l, u]$ found directly. If the two intervals for $\theta(F)$ are identical, then the property of 'transformation invariance' holds; if not, it makes sense to ask "which of the two intervals is more accurate?", in which case one is led to search for an 'optimal' transformation $g(\cdot)$ to use in connection with the construction of confidence intervals.

In some isolated cases, e.g., Fisher's hyperbolic tangent transformation for the correlation coefficient (cf. [20] p. 54 and p. 163, and [58]), a transformation is available in the literature that approximately 'normalizes' and 'variance stabilizes' the estimator $T(\mathbf{X})$; in other words, the estimator $g(T)$ has a distribution that is closer to being Gaussian than the distribution of $T(\mathbf{X})$, and the variance of $g(T)$ does not depend on the parameter $\theta(F)$, at least not significantly. As

a consequence, such a transformation is 'optimal' to use in connection with the construction of confidence intervals based on the normal approximation of equation (3).

In most cases however, it may not be possible to simultaneously normalize and variance stabilize the estimator $T(\mathbf{X})$ by a single transformation. As it turns out, the 'optimal' transformation associated with constructing bootstrap-t confidence intervals should primarily achieve variance stabilization. Now if $Var_F(T)$ were *known* as a function of $\theta(F)$, then an approximate variance stabilizing transformation $g(\cdot)$ could be found by the δ-method (cf. [37], [20]). The problem of course is that $Var_F(T)$, $\theta(F)$ –as well as the functional relationship between the two– are generally unknown!

Nonetheless, an approximate 'optimal' transformation for variance stabilization can be computed using an iterated bootstrap –much like the iterated bootstrap described in the previous Subsection on studentization– to calculate estimates of $Var_F(T)$ from each resample; details can be found in [20] (p. 163), and in [58]. It should be noted that if an iterated bootstrap is carried out to calculate the variance stabilizing transformation, then there is no need to do another iterated bootstrap to get the bootstrap-t confidence interval. In other words, there is no need for the studentization any more since the variance can be considered constant, and a bootstrap confidence interval for $g(\theta)$ based on the root method of equation (18) would be obtained and then inverted (using g^{-1}) to give a good bootstrap confidence interval for $\theta(F)$.

2. Subsampling and the jackknife

While one reason for the success of the bootstrap is its widespread applicability, there are certainly situations where the bootstrap is *not* applicable; for example, as was briefly mentioned in Subsection 1.8, in the case where the statistic $T(\mathbf{X})$ is linear, i.e., of the sample mean type, the validity of the bootstrap crucially hinges on whether the statistic is asymptotically normal or not. As a matter of fact, a huge statistical literature on the bootstrap has accumulated since Efron's [14] pioneering paper, with main focus to show the applicability of the bootstrap in many different settings; [8] and [3] are two very important early papers in this connection. Note also that the jackknife of Quenouille [48] and Tukey [55] is a technique closely related to the bootstrap, and is generally thought of as the precursor of bootstrap. Nevertheless, although the jackknife has been around longer than the bootstrap, to motivate the subsequent discussion on the jackknife let us now discuss some examples where the bootstrap does *not* work.

2.1 The bootstrap does not always work! Examples of failure of the bootstrap have

been provided from the very beginning of the development of bootstrap methodology. One of the earliest counterexamples [8] concerns the case where the data $X_1, ..., X_N$ are i.i.d. $Uniform$ $(0, \theta)$, and $T(\mathbf{X}) = \max\{X_1, ..., X_N\}$. Another interesting example of bootstrap failure is given [1] when $T(\mathbf{X})$ is the familiar sample mean $\bar{X}$ of i.i.d. data $X_1, ..., X_N$ but where the data come from a distribution being in the domain of attraction of a (nonnormal) stable law. For example, if $X_1, ..., X_N$ are i.i.d. from a standard *Cauchy* distribution[9] then the bootstrap will behave erratically even if the sample size is huge; more examples of bootstrap failure are given in [41], [43], and [44].

The reason the bootstrap may fail can be pin-pointed to the fact that the resamples are not exactly generated from F (as the original sample was), but are generated from $\hat{F}$ instead. [10] The question now is: "*can* we find samples/resamples exactly generated from F, and -if yes- where?". If we do not insist that we find (re)samples of size N, and we are content with finding (re)samples of size b (with $b < N$), then the answer is that we *can* indeed find (re)samples of size b exactly generated from F simply *by looking at different subsets of our original sample* $\mathbf{X}$. But looking at different subsets of our original sample amounts to sampling *without* replacement from the observations $X_1, \ldots, X_N$ to get (re)samples (now called *subsamples*) of size b; this leads us to subsampling and the jackknife.

2.2 The jackknife idea. Consider sampling without replacement from the observations $X_1, \ldots, X_N$, to get a *subsample* of size b, where of course $b < N$. If $b = N - 1$, this is exactly the original jackknife of Quenouille [48] and Tukey [55], and there are only N possible different subsamples (and their permutations); cf. [14], [16], and [20] for details and extensive list of references. Since these subsamples are all equally probable under the sampling-without-replacement scheme, formulas much like (15), (16), (19), and (20) can be constructed to estimate bias, variance, and distribution of the statistic $T(\mathbf{X})$; these will be given in a more general form in what follows.

The jackknife with $b = N - 1$ is definitely less computer-intensive than the bootstrap which was perhaps on of the reasons its development preceded that of the bootstrap; see e.g. chapter 11 of [20]. Nevertheless, one can take an arbitrary b, not necessarily equal to $N - 1$, yielding the so-called delete-d jackknife, where $d = N - b$; see e.g. [51]. Observe that the number of possible

[9] Of course, here $\bar{X}$ is not asymptotically normal so the Central Limit Theorem breaks down as well; $\bar{X}$ is itself Cauchy-distributed.

[10] Although $\hat{F}$ is a consistent estimator of the true distribution F it may still miss some feature of F that crucially affects the distribution of $T(\mathbf{X})$; another way of explaining bootstrap failure is to note that the expressions $Dist_{T,F}(x)$, $Dist_{T-\theta,F}(x)$, etc. might not be (uniformly) continuous in F, and consequently the 'plug-in' principle may fail.

subsamples now rises to $\frac{N!}{b!(N-b)!}$ (and their distinct permutations[11]), and again a Monte Carlo method could be employed to randomly chose a smaller number, say B, among these subsamples to be included in the jackknife procedure. Now it is a very encouraging finding that, if both b and N are large, but with b being small with respect to N, (i.e., if $b \to \infty$ but $b/N \to 0$), subsampling (i.e., sampling without replacement) generally remedies the failure of the bootstrap in most cases (including all the examples mentioned in Subsection 2.1); cf. [43].

In some sense, subsampling can be thought to be even more intuitive than the bootstrap, because the subsamples are actually samples (of smaller size) from the *true* distribution F, whereas the bootstrap resamples are samples from an estimator of F. As can be shown, distribution estimates based on subsampling are valid in a wider range of situations than their resampling (i.e., bootstrap) analogs, [12] even in cases where the statistic $T(\mathbf{X})$ is *not* asymptotically normal; however, they do not possess the property of higher order accuracy, and this is essentially due to the fact that the subsampling size is b and not N.

This difference between the subsample size and the original sample size has an additional consequence, namely that a re-scaling is in order in computing the subsampling distribution estimator. Suppose that the variance of $T(\mathbf{X})$ is approximately c^2/τ_N^2, for large N, where c is some constant, and τ_N is a function [13] of the sample size N. It is intuitively true therefore that the variance of T calculated from a sample of size b would be approximately c^2/τ_b^2, provided that b is large too; here the need for a re-scaling becomes apparent. The subsampling procedure can finally be summarized as follows:

[11] Statistics from i.i.d. data are usually symmetric in their arguments, i.e., if $\tilde{\mathbf{X}}$ is a permutation of $\mathbf{X}$, then $T(\tilde{\mathbf{X}})$=$T(\mathbf{X})$. For example, if $T(\mathbf{X})$=$\theta(\hat{F})$, then $T(\mathbf{X})$ is certainly symmetric, and thus insensitive to different permutations of the same dataset; in other words, the statistic T re-evaluated over all possible subsample of size b will take on (at most) $\frac{N!}{b!(N-b)!}$ different values.

[12] It should be noted here that sampling *with* or *without* replacement from the original sample population $\{X_1, \ldots, X_N\}$ would make no difference in practice if b is *very* small with respect to N (i.e., if $b/\sqrt{N} \to 0$); in other words, there is no practical difference between resampling and subsampling as long as the resample/subsample size b is very small (i.e., $b/\sqrt{N} \approx 0$). This phenomenon is analogous to the binomial approximation to the hypergeometric in sampling from a finite (but large) population. Therefore, the bootstrap (i.e., sampling with replacement) with smaller resample size b will also work in many cases where the bootstrap with resample size N fails, but will lack the property of higher order accuracy possessed by the standard bootstrap that has resample size N. To avoid possible confusion, the bootstrap with resample size b will not be discussed any further here; cf. [41], [43], [44], [6], and [9] for more details.

[13] The functional form of τ_N is problem-specific and should be calculated for the problem at hand; typically, $\tau_N = N^a$, with a being a constant in (0,1]. In regular cases, e.g., if $T(\mathbf{X})$ is the sample mean, sample median, sample variance, etc., we have that $a = 1/2$ and $\tau_N = \sqrt{N}$. If τ_N is difficult to calculate, or if its form (say, the exponent a) depends on the unknown distribution F, then it is necessary to estimate τ_N from the data at hand which can be achieved by a preliminary round of subsampling [7].

- Randomly choose B subsamples $\mathbf{X}^{\star(1)}, \ldots, \mathbf{X}^{\star(B)}$ among all the possible subsamples of size b of the sample population $\{X_1, \ldots, X_N\}$. Suppose the ith subsample is $\mathbf{X}^{\star(i)} = (X_1^{\star(i)}, \ldots, X_b^{\star(i)})$; the final step now is to evaluate the statistic T over each of the chosen subsamples, creating the pseudo-replications $T(\mathbf{X}^{\star(1)}), \ldots, T(\mathbf{X}^{\star(B)})$.

2.3 Confidence intervals based on subsampling. The subsampling estimates of $Bias_F(T)$, $Var_F(T)$, $Dist_{T,F}(x)$, and $Dist_{T-\theta,F}(x)$ are $Bias^\star(T)$, $Var^\star(T)$, $Dist^\star_{T,F}(x)$, and $Dist^\star_{T-\theta,F}(x)$ respectively which are presented below:

$$Bias^\star(T) \simeq \frac{\tau_r}{\tau_N}\left(\frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{\star(i)}) - T(\mathbf{X})\right) \tag{30}$$

$$Var^\star(T) \simeq \frac{\tau_r^2}{\tau_N^2}\left(\frac{1}{B}\sum_{i=1}^{B} T^2(\mathbf{X}^{\star(i)}) - [\frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{\star(i)})]^2\right) \tag{31}$$

$$Dist^\star_{T,F}(x) \simeq \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}(T(\mathbf{X}^{\star(i)}) \leq x\frac{\tau_N}{\tau_r}) = \frac{1}{B}(\#T(\mathbf{X}^{\star(i)}) \leq x\frac{\tau_N}{\tau_r}) \tag{32}$$

and

$$Dist^\star_{T-\theta,F}(x) \simeq \frac{1}{B}(\#T(\mathbf{X}^{\star(i)}) \leq x\frac{\tau_N}{\tau_r} + T(\mathbf{X})), \tag{33}$$

where $r = bN/(N-b)$; note that if $B = \frac{N!}{b!(N-b)!}$ and Monte Carlo randomization is not used, i.e., *all* possible subsamples are taken into account, the approximation signs ($\simeq$) above can be replaced by equality signs.

The reason we have τ_r instead of the more intuitive τ_b in (30), (31), (32), and (33) is that, although the variance of T calculated from an i.i.d. sample of size b is approximately c^2/τ_b^2, i.i.d. samples presuppose an infinite underlying population; in other words, i.i.d. samples are taken *with* replacement. Our subsamples $\mathbf{X}^{\star(1)}, \ldots, \mathbf{X}^{\star(B)}$ are size b samples taken *without* replacement from a *finite* population of size N. Therefore, the variance of T calculated from one of our subsamples is approximately c^2/τ_r^2, and not c^2/τ_b^2; the ratio τ_r^2/τ_b^2 is the so-called *finite population correction*, which notably becomes close to one *provided b is much smaller than N.*

To briefly sum-up the existing results in the literature on subsampling, note that the estimates proposed in equations (30), (31), (32), and (33) are accurate provided the sample size N is large, and that one of the following three conditions is met:

(a) $T(\mathbf{X})$ is a linear statistic (i.e., $T(\mathbf{X}) = \theta(\hat{F})$, with $\theta(.)$ being linear), and the population distribution F is *known* to be normal (with some unknown mean and variance); for example, $T(\mathbf{X})$ may be the sample mean, or maybe a trimmed mean (but not the sample median!). In this case, the ordinary jackknife (with $b = N-1$) would work [20].

(b) The estimator is not necessarily that simple, but it is asymptotically normal, and satisfies the 'regularity' condition that $\tau_N = \sqrt{N}$; for example, $T(\mathbf{X})$ may be the sample median. In this case we would have to choose a b such that both b and $N-b$ are large, e.g., $b = \lfloor kN \rfloor$, where k is a constant in $(0,1)$, and $\lfloor . \rfloor$ denotes the integer part; see [51].

(c) The estimator is arbitrarily complex, not necessarily asymptotically normal, and τ_N is not necessarily $\sqrt{N}$; here we would have to choose a b such that b is large, but b/N is small, e.g. $b = \lfloor N^k \rfloor$, where k is a constant in $(0,1)$. Note that in this case N and $N-b$ will be of the same approximate magnitude, thus $r \approx b$; therefore, equations (30), (31), (32), and (33) will be valid with τ_b used instead of τ_r, which is perhaps more intuitive; see [43] for more details on this general case.

Under one of conditions (a), (b) or (c) above, the approximations proposed in equations (30), (31), (32), and (33) are accurate, and can be used for the construction of approximate confidence intervals for $\theta(F)$; as is typically the case, the approximations will be good if the sample size N is appropriately large. Note that in the usual case where we do not know the form of the population distribution F, we have to use the settings of conditions (b) or (c) above, i.e., to assume that both b and $N-b$ are large. Now if the estimator $T(\mathbf{X})$ is known to be asymptotically normal (see our Subsection 1.1), then a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ can be given by

$$[T(\mathbf{X}) - Bias^\star(T) - z\sqrt{Var^\star(T)}, T(\mathbf{X}) - Bias^\star(T) + z\sqrt{Var^\star(T)}] \tag{34}$$

where $z = z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of the standard normal distribution as in (4). If $T(\mathbf{X})$ is not asymptotically normal, then an equal-tailed $(1-\alpha)100\%$ confidence interval for $\theta(F)$ based on subsampling could be constructed similarly to the interval (18), i.e., it would be given by

$$[T(\mathbf{X}) - q^\star(1-\alpha/2), T(\mathbf{X}) - q^\star(\alpha/2)], \tag{35}$$

where $q^\star(\alpha/2)$ and $q^\star(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist^\star_{T,F,\theta}(x)$ distribution respectively. Note that if both confidence intervals (34) and (35) are valid, i.e., if $T(\mathbf{X})$ is asymptotically normal, the interval (34) would be considered to be the one that is more accurate; that is, the coverage probability of (34) would be closer to the desired value of $1-\alpha$. In other words, whereas the (studentized) bootstrap 'beats' the normal approximation (if a normal approximation is available), the jackknife does not. Nevertheless, subsampling (and the interval (35) in particular) is more widely applicable than either the bootstrap, or the normal approximation, as it is valid in the general setting of condition (c) above; note that, as discussed in Subsection 1.8, the bootstrap is typically valid under the more restrictive conditions (a) or (b).

3. Non-i.i.d. data: complicated data structures

What has been discussed so far hinges on the assumption that the data $\mathbf{X} = (X_1, \ldots, X_N)$ represent an i.i.d. sample from a population with (unknown) distribution F. Nevertheless, the assumption of i.i.d. data can break down, either because the data are not independent, or because they are not identically distributed, or both; we now discuss what can be done to circumvent this difficulty, and describe procedures that are are valid even if the i.i.d. assumption is somehow violated.

3.1 Data that are not identically distributed: the regression example. To fix ideas, let us consider the simplest example. Suppose the X_i, $i = 1, ..., N$, are observations from the straight-line regression model $X_i = \gamma + \beta Y_i + \epsilon_i$, where the ϵ_i's are assumed to be i.i.d. with mean zero, and Y_i, $i = 1, ..., N$, are known (nonrandom) design points.

Let $\hat{\gamma}, \hat{\beta}$ denote the least squares estimates of the intercept γ and of the slope β respectively, and suppose we want to construct a confidence interval for one of the two parameters, say β. It is well-known that, regardless of whether the 'errors' ϵ_i, $i = 1, ..., N$, are normally distributed or not, $\hat{\beta}$ is a reasonable estimator of β; as a matter of fact, typically $\hat{\beta}$ will be consistent for β, i.e., for large sample size N, $\hat{\beta}$ will be close to the true value β with high probability.

Nevertheless, the standard textbook confidence interval for β is based on the assumption that the ϵ_i's are normal; if the normality assumption is questionable,[14] then an alternative method of constructing the confidence interval must be found. The bootstrap may offer this well-needed alternative. However, note that the X_i's are *not* i.i.d.; in particular, $EX_i = Y_i$ which varies with $i = 1, ..., N$. In other words, although the X_i's are independent, they are not identically distributed; thus, naive resampling of the X_i's can not be applied here.

To actually apply the bootstrap in this setting observe that, whereas in the previous Sections the data X_i, $i = 1, ..., N$, were i.i.d. with unknown distribution F, here it is the errors ϵ_i, $i = 1, ..., N$, that are i.i.d. with unknown distribution which we may denote by $G(\cdot)$. Although the errors ϵ_i are not directly observable, note that $\hat{\beta}$ will be close to the true value β (and similarly $\hat{\gamma}$ will be close to the true value γ), and thus $\hat{\gamma} + \hat{\beta} Y_i$ will be close to $\gamma + \beta Y_i$, for any $i = 1, ..., N$. It follows that the $e_i \equiv X_i - (\hat{\gamma} + \hat{\beta} Y_i)$, $i = 1, ..., N$, i.e., the residuals from the least squares fit, will be good approximations to the unobservable i.i.d. errors ϵ_i.

[14] For example, a histogram of the shortly-to-be-defined residuals e_i, $i = 1, ..., N$, may exhibit evidence of nonnormality, e.g., pronounced skewness.

So we may treat the residuals e_i as being an i.i.d. sample from distribution $G(\cdot)$, *provided* the e_i's have mean zero; note that although $G(\cdot)$ is unknown, we do know that $G(\cdot)$ is a distribution with zero mean. Thus we are led to define the mean-corrected [15] residuals $\hat{\epsilon}_i = e_i - N^{-1}\sum_{k=1}^{N} e_k$. We can now invoke the bootstrap principle and treat the $\hat{\epsilon}_i$'s as if they represented an i.i.d. sample from distribution $G(\cdot)$. Let $\hat{G}(x)$ be the empirical distribution of the $\hat{\epsilon}_i$'s, i.e., let $\hat{G}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{\epsilon}_i \leq x) = \frac{1}{N}(\#\hat{\epsilon}_i \leq x)$. The bootstrap resampling in this case can be done as follows:

- Draw B i.i.d. samples $\mathbf{E}^{*(1)}, \ldots, \mathbf{E}^{*(B)}$ (each of size N) from the sample population consisting of the mean-corrected residuals $\{\hat{\epsilon}_1, ..., \hat{\epsilon}_N\}$; note that to form the kth resample $\mathbf{E}^{*(k)} = (\hat{\epsilon}_1^{*(k)}, \ldots, \hat{\epsilon}_N^{*(k)})$, we sample with replacement from the set $\{\hat{\epsilon}_1, ..., \hat{\epsilon}_N\}$, or, in other words, we take an i.i.d. sample of size N from a population with distribution $\hat{G}$. Use the kth resample $\mathbf{E}^{*(k)}$ to generate pseudo-data $X_i^{*(k)} = \hat{\gamma} + \hat{\beta}Y_i + \hat{\epsilon}_i^{*(k)}$, $i = 1, ..., N$, where the $\hat{\gamma}$, $\hat{\beta}$, are the previously (and once and for all) calculated least squares estimators. Now apply least squares estimation to the kth pseudo-dataset to obtain the estimator $\hat{\beta}^{*(k)}$; repeat this procedure for each of the kth resamples, $k = 1, ..., B$.

Having performed the above Monte Carlo experiment, we are now in a position to formulate estimates of the bias and variance of $\hat{\beta}$ similar to equations (15), i.e.,

$$Bias^*(\hat{\beta}) \simeq \frac{1}{B}\sum_{i=1}^{B} \hat{\beta}^{*(k)} - \hat{\beta} \tag{36}$$

$$Var^*(\hat{\beta}) \simeq \frac{1}{B}\sum_{i=1}^{B} (\hat{\beta}^{*(k)})^2 - [\frac{1}{B}\sum_{i=1}^{B} \hat{\beta}^{*(k)}]^2 \tag{37}$$

and an equal-tailed root $(1-\alpha)100\%$ bootstrap confidence interval for β similar to the one in equation (18), i.e.,

$$[\hat{\beta} - g^*(1-\alpha/2), \hat{\beta} - g^*(\alpha/2)]; \tag{38}$$

here $g^*(\alpha/2)$ and $g^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution $Dist_{\hat{\beta},\hat{G},\beta}(x) \equiv B^{-1}\sum_{k=1}^{B} \mathbf{1}(\hat{\beta}^{*(k)} - \hat{\beta} \leq x) = B^{-1}(\#\hat{\beta}^{*(k)} \leq x + \hat{\beta})$.

Note that different confidence interval constructions are possible here as well, [16] in analogy to what was discussed for the i.i.d. bootstrap in Section 1. As a matter of fact, using the trick

[15] As a matter of fact, the inclusion of the unknown (and to be estimated) intercept parameter γ in the regression model is sufficient to guarantee that the original residuals e_i do have mean zero, and thus $\hat{\epsilon}_i = e_i$ in this case. Nevertheless, it is important to point out that if, for some reason (e.g., a more complicated regression function of the type $X_i = \Gamma(Y_i) + \epsilon_i$), the residuals from the fitted model are not forced to have mean zero, then the corresponding bootstrap inferences will not be valid.

[16] At this point it should also be mentioned that the regression model could be 'bootstrapped' by bootstrapping the pairs (X_i, Y_i), $i = 1, ..., N$, rather than the residuals; see [20] (p. 113) for a discussion. Note, however, that 'bootstrapping pairs' implies in effect that the Y_i points are *not* at the choice of the person designing the experiment, but that they are random.

of reducing the non-i.i.d. situation to an i.i.d. situation (by looking at the residuals) permitted us to use the bootstrap methodology for i.i.d. data with almost no alterations. This is actually a general technique, applicable in all settings where the problem can be reduced to i.i.d. (or almost i.i.d.) residuals.

In general, the regression model might be more complicated, e.g., the relation of X_i to Y_i might be nonlinear, and described by $X_i = \Gamma(Y_i) + \epsilon_i$, where the ϵ_i's are i.i.d. with mean zero, the Y_i's are known and nonrandom, and $\Gamma(.)$ is an unknown function; perhaps Γ is known, except for some parameters associated with it, e.g., $\Gamma(y) = e^{\gamma+\beta Y_i}$, with γ, β unknown as before. If an estimator $\hat{\Gamma}(.)$ of $\Gamma(.)$ can be constructed from the data, the residuals $X_i - \hat{\Gamma}(Y_i)$ should be almost i.i.d. (and can be centered so that they have mean zero), so the i.i.d. bootstrap methodology described above applies immediately.

It should also be pointed out that in regression analysis, besides estimation of the unknown parameters γ and β (or $\Gamma(.)$ in general), the practitioners are typically interested in predicting the value of X corresponding to a future Y-point, and attaching a measure of accuracy to their prediction. The bootstrap can help out here as well, and can be successfully employed for the construction of *prediction* (rather than confidence) intervals; see, e.g., [20] (p. 247) and [31].

3.2 Data that are not independent: the autoregressive time series example. Another way to relax the assumption of i.i.d. data is to assume that the data are identically distributed but not independent; this is the essence of the stationarity assumption. In particular, let $X_1, X_2, ...$ be a sequence of random variables; the sequence is called *strictly stationary* (or just stationary) if the joint distribution of the random vector $(X_1, X_2, ..., X_n)$ is identical to the joint distribution of the random vector $(X_{m+1}, X_{m+2}, ..., X_{m+n})$, for any positive integers m, n.

The simplest example of a strictly stationary sequence is given by the autoregression model (of order one) that satisfies the recursion $X_i = \beta X_{i-1} + \epsilon_i$, where the ϵ_i's are i.i.d. with mean zero; for simplicity, let us assume that the X_i's have mean zero, so no constant term[17] is included in the right-hand-side of the above recursion. It is also usually assumed that $|\beta| < 1$, in which case we may consider that the X_i's were obtained in a "causal" fashion, by letting $i = 1, ..., N$ in the defining recursion, and with some proper choice of X_0.

It is apparent that since the parameters in the model (in this case it is just β) can be estimated

[17] Note however that, although the theoretical mean is zero, the sample mean of the X-dataset, i.e., $N^{-1}\sum_{i=1}^{N} X_i$, will *not* be identically zero. Working with the mean-corrected X_i's, i.e., defining $\tilde{X}_i = X_i - N^{-1}\sum_{i=1}^{N} X_i$ and working with the model $\tilde{X}_i = \beta\tilde{X}_{i-1} + \epsilon_i$ instead, is recommendable in practice; it also has the convenient side-effect of forcing the residuals from the fitted model $\tilde{X}_i = \beta\tilde{X}_{i-1} + \epsilon_i$ to have mean zero, so that no re-centering would be required (see [58] as well).

consistently, the i.i.d. errors ϵ_i, $i = 1, ..., N$, can be approximately recaptured; in other words, the stationary data problem at hand can be reduced to an (approximate) i.i.d. problem, in the same spirit as in the regression example of Subsection 3.1. As a matter of fact, by making the formal identification $Y_i \equiv X_{i-1}$, for $i = 1, ..., N$, the bootstrap algorithm described in Subsection 3.1 applies *verbatim* here as well, although the Y_i are no longer nonrandom; for more details on the bootstrap for linear or nonlinear autoregressive time series models (including cases where the order of autoregression is infinite) see [10], [28], [31], [38], [58].

3.3 Data that are not too dependent: weakly dependent observations. Suppose that no plausible model (such as the autoregression of Subsection 3.2) is available for the probability mechanism generating our stationary observations $X_1, \ldots, X_N$; in this case, the problem must be approached in a nonparametric fashion. Nonetheless, in order to have consistent estimation of θ by $T(\mathbf{X})$, i.e., in order to be able to say that 'the more data available, the more accurate our inference is', the observations should not be too strongly dependent; for example, in the extremely dependent case where $X_j = X_1$, for $j = 1, 2, ..., N$, obtaining more data (i.e., increasing N) does not tell us something we do not already know by looking at X_1 alone.

So an assumption of weak dependence must be made in order that consistent estimation is possible. One such assumption is m-dependence: the stationary sequence $X_1, X_2, \ldots$ is called m-dependent if, for some integer m, the set of random variables $(X_1, X_2, ..., X_n)$ is independent of the set of random variables $(X_{n+k+1}, X_{n+k+2}, ...X_{2n+k})$ for any n and any $k \geq m$; thus, independence can be thought of as 0-dependence. Another weak dependence assumption is *strong mixing*: although the precise definition is a bit technical [39], [43], the intuitive idea is that observations far apart (in time) should be almost independent; more carefully, a stationary sequence $X_1, X_2, \ldots$ is strong mixing if the set of random variables $(X_1, X_2, ..., X_n)$ is *approximately* independent of the set of random variables $(X_{n+k+1}, X_{n+k+2}, ..., X_{2n+k})$, for any n, as long as k is large enough. Note that an m-dependent sequence is definitely strong mixing; just let $k \geq m$ in the above.

3.4 Subsampling weakly dependent observations. Let $X_1, X_2, \ldots$ be a strong mixing stationary sequence of random variables, and suppose our data consist of the stretch $X_1, X_2, ..., X_N$. Note that the order of the observations in our sample $X_1, X_2, ..., X_N$ is important now that the X_i's are serially dependent, whereas it was not important in the case the X_i's were independent.

So consider the $N - b + 1$ subsamples characterized by the property that each contains b *consecutive* observations from the original sample $X_1, \ldots, X_N$; in this sense, *the time order of the observations is maintained within the subsamples.* For example, the ith subsample $\mathbf{X}^{\star(i)}$

would consist of the *block* of consecutive observations $X_i, ..., X_{i+b-1}$, where now i is a positive integer with $i \leq N - b + 1$. Note that now the number of subsamples consisting of b consecutive data is $N - b + 1$ which is rather small compared to $\frac{N!}{b!(N-b)!}$; thus, Monte Carlo randomization typically will not be needed and we would choose $B = N - b + 1$ subsamples (i.e., all of them) to be included in the subsampling procedure as outlined in Subsection 2.2. In other words, the statistic T would be evaluated over each of the $B = N - b + 1$ subsamples creating the pseudo-replications $T(\mathbf{X}^{\star(1)}), \ldots, T(\mathbf{X}^{\star(B)})$.

Interestingly enough, this modification (i.e., looking only at subsamples containing consecutive X_i's) is sufficient to make the subsampling methodology work in this case where the observations are stationary (and weakly dependent); see [43] for more details. Note, however, that the presence of the dependence here forces us to function in the general setting of condition (c) of Subsection 2.3, i.e., we are $forced$ to choose a b such that b is large, but b/N is small, e.g. $b = \lfloor N^k \rfloor$, where k is a constant in $(0, 1)$; standard choices of b would be $b = \lfloor N^{1/2} \rfloor$ or $b = \lfloor N^{1/3} \rfloor$. With such a choice for b, equations (30), (31), (32), and (33) would apply here *verbatim* and they would give accurate estimators of bias, variance, and distribution if the sample size N is large enough. Consequently, (35) would give a valid $(1 - \alpha)100\%$ confidence interval for $\theta(F)$, whereas (34) would also give a valid $(1 - \alpha)100\%$ confidence interval for $\theta(F)$ provided the estimator $T(\mathbf{X})$ is known to be asymptotically normal.

3.5 Resampling weakly dependent observations. Again let us assume that our data consist of the stretch $X_1, X_2, ..., X_N$ from the strong mixing stationary sequence $X_1, X_2, \ldots$. Unlike the bootstrap for i.i.d. data, a bootstrap for stationary observations would have to somehow maintain the time order of the observations as was done in the subsampling case in Subsection 3.4. This is the essense of the 'moving blocks' bootstrap [29], [35]; see also [39], [42] for closely related proposals for bootstrapping dependent data.

Consider the set $\mathcal{S} = \{\mathbf{X}^{\star(1)}, \mathbf{X}^{\star(2)}, ..., \mathbf{X}^{\star(N-b+1)}\}$, where $\mathbf{X}^{\star(i)} = (X_i, ..., X_{i+b-1})$ is the ith subsample defined in Subsection 3.4; so $\mathcal{S}$ is a set of subsamples. As in Subsection 3.4, here as well we require that b is large, but b/N is small, e.g. $b = \lfloor N^k \rfloor$, where k is a constant in $(0, 1)$. Let $K = \lfloor N/b \rfloor$, and let $L = bK$; thus, $L = N$ if N is divisible by b, whereas L will at least approximate N if N is not exactly divisible by b (because N/b is assumed to be large).

The 'moving blocks' bootstrap can be described as follows;

- Take a random sample of size K with replacement from the set $\mathcal{S}$, i.e., randomly choose subsamples $\mathbf{X}^{**(1)}, \ldots, \mathbf{X}^{**(K)}$; concatenate the observations found in $\mathbf{X}^{**(1)}, \ldots, \mathbf{X}^{**(K)}$ into a series of $L = bK$ observations denoted by $\mathbf{Y}^{**(1)}$. Take another random sample of size

K with replacement from the set $\mathcal{S}$, and store it in $\mathbf{Y}^{**(2)}$. In the same manner, generate $\mathbf{Y}^{**(i)}$, for $i = 3, 4, ..., B$. Now evaluate the statistic T over each of the $\mathbf{Y}^{**(i)}$, for $i = 1, 2, 3, ..., B$, bootstrap pseudo-series to get the pseudo-replications $T(\mathbf{Y}^{**(1)}), ..., T(\mathbf{Y}^{**(B)})$.

Note that the requirement that b is large is only pertinent if the data are indeed dependent; if the date are i.i.d., b can be taken to equal 1, and the 'moving blocks' bootstrap actually reduces to the standard i.i.d. bootstrap described in Section 1. If the data are just suspected to be serially dependent, then the 'moving blocks' bootstrap (with large b as opposed to b=1) can be employed in order to be on the safe side.

The 'moving blocks' bootstrap estimates of $Bias_F(T)$, $Var_F(T)$, $Dist_{T,F}(x)$, and $Dist_{T-\theta,F}(x)$ are $Bias^{**}(T)$, $Var^{**}(T)$, $Dist^{**}_{T,F}(x)$, and $Dist^{**}_{T-\theta,F}(x)$ respectively which are presented below:

$$Bias^{**}(T) \simeq \left(\frac{1}{B} \sum_{i=1}^{B} T(\mathbf{Y}^{**(i)}) - T(\mathbf{X}) \right) \tag{39}$$

$$Var^{**}(T) \simeq \left(\frac{1}{B} \sum_{i=1}^{B} T^2(\mathbf{Y}^{**(i)}) - [\frac{1}{B} \sum_{i=1}^{B} T(\mathbf{Y}^{**(i)})]^2 \right) \tag{40}$$

$$Dist^{**}_{T,F}(x) \simeq \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}(T(\mathbf{Y}^{**(i)}) \leq x) = \frac{1}{B}(\#T(\mathbf{Y}^{**(i)}) \leq x) \tag{41}$$

and

$$Dist^{**}_{T-\theta,F}(x) \simeq \frac{1}{B}(\#T(\mathbf{Y}^{**(i)}) \leq x + T(\mathbf{X})). \tag{42}$$

Similarly to Subsection 3.4, an equal-tailed root $(1-\alpha)100\%$ confidence interval for $\theta(F)$ based on the 'moving blocks' bootstrap would be given by

$$[T(\mathbf{X}) - q^{**}(1-\alpha/2), T(\mathbf{X}) - q^{**}(\alpha/2)], \tag{43}$$

where $q^{**}(\alpha/2)$ and $q^{**}(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist^{**}_{T,F,\theta}(x)$ distribution respectively.

Note that, as opposed to the subsampling method of Subsection 3.4, no rescaling is needed for the 'moving blocks' bootstrap (as it was not needed in the i.i.d. bootstrap as well); this is because $L \approx N$, and therefore $\tau_L/\tau_N \simeq 1$. In addition, the 'moving blocks' bootstrap shares with the i.i.d. bootstrap the property of higher order accuracy as was discussed in Subsection 1.8. In other words, if $T(\mathbf{X})$ is asymptotically normal, the 'moving blocks' bootstrap can be applied to an appropriately 'studentized' version of $T(\mathbf{X})$ to yield confidence intervals that are more accurate than the intervals obtained from the normal approximation; cf. [30] and [23]. Nevertheless, there are situations where the 'moving blocks' bootstrap would not be applicable, and subsampling would provide the only solution; for example, a requirement for the 'moving

blocks' bootstrap to 'work' is that $\tau_N = \sqrt{N}$, i.e., that the variance of $T(\mathbf{X})$ is (for large N) approximately proportional to $1/N$, *and* that $T(\mathbf{X})$ is indeed asymptotically normal [49].

3.6 A 'difficult' example: nonparametric confidence intervals for the spectrum. As before, our data consist of the stretch $X_1, X_2, ..., X_N$ from the strong mixing stationary sequence $X_1, X_2, \ldots$ which for simplicity we now assume to have mean zero, i.e., $EX_n = 0$, for any n. Let $R(s) = EX_0X_{|s|}$ denote the autocovariance at 'lag' s; as a consequence of strong mixing, it can be shown that $R(s) \to 0$ as $|s| \to \infty$. If we assume in addition that $R(s) \to 0$ fast enough such that $\sum_{s=-\infty}^{\infty} |R(s)| < \infty$, then we can define the spectral density function $f(w) = \sum_{s=-\infty}^{\infty} R(s)e^{-jsw}$; here w is a point in $[0, 2\pi]$, and j is the imaginary unit, i.e., $\sqrt{-1}$.

Suppose that the problem at hand is interval estimation of $f(w_0)$, where w_0 is a point of interest in $[0, 2\pi]$; thus, the unknown parameter of interest is $\theta = f(w_0)$. Suppose also that for this purpose we decide to employ Bartlett's spectral density estimator given by $\hat{f}(w) = \sum_{s=-\infty}^{\infty} \hat{R}(s)\lambda_M(s)e^{-jws}$, where $\lambda_M(s)$ is Bartlett's kernel defined by

$$\lambda_M(s) = \begin{cases} 1 - \frac{|s|}{M} & \text{if } |s| \le M \\ 0 & \text{for } |s| > M, \end{cases}$$

and $\hat{R}(s)$ is the sample autocovariance at lag s given by

$$\hat{R}(s) = \begin{cases} N^{-1} \sum_{i=1}^{N-|s|} X_i X_{i+|s|} & \text{if } 0 \le |s| \le N \\ 0 & \text{if } |s| > N. \end{cases}$$

It is well known [47] that, under some regularity conditions, $\hat{f}(w)$ is asymptotically normal, and that $Var(\hat{f}(w)) \approx \frac{8\pi^2 M}{3N} f^2(w)(1 + \eta(w))$, if N is large, where $\eta(w) = 0$ if $w \neq 0 (mod \pi)$ and $\eta(w) = 1$ if $w = 0 (mod \pi)$. It is also well known that to minimize the Mean Squared Error of the estimator $\hat{f}(w)$ we should choose M to be approximately proportional to $N^{1/3}$; this is because the bias of $\hat{f}(w)$ is approximately (for large N) proportional to $1/M$. So let us choose $M = AN^{1/3}$, where A is some positive constant; thus

$$T(\mathbf{X}) \equiv \hat{f}(w_0) = \sum_{s=-\infty}^{\infty} \hat{R}(s)\lambda_{AN^{1/3}}(s)e^{-jw_0 s}. \tag{44}$$

With this choice of M the variance of $\hat{f}(w)$ becomes approximately (for large N) proportional to $N^{-2/3}$; in other words, *the variance of $T(\mathbf{X})$ is (for large N) approximately proportional to $N^{-2/3}$, and $\tau_N = N^{1/3}$.*

Since the variance of $T(\mathbf{X})$ is *not* approximately proportional to $1/N$, it should not be surprising that the 'moving blocks' bootstrap does not apply here. A generalization of the 'moving blocks' bootstrap (the so-called 'blocks of blocks' bootstrap) was introduced in [45] in order to

handle this 'difficult' example; see also [21] for a different approach –familiar to us from the regression example– based on bootstrapping residuals. Nonetheless, the subsampling methodology as described in Subsection 3.4 *does* apply, provided we choose b such that b is large, but b/N is small, e.g., $b = [N^k]$, for some constant k in $(0, 1)$; see condition (c) in Subsection 3.4. To elaborate, let the ith subsample $\mathbf{X}^{\star(i)}$ consist of the observations $X_i, ..., X_{i+b-1}$; applying the statistic T on the subsample $\mathbf{X}^{\star(i)}$ amounts to letting

$$T(\mathbf{X}^{\star(i)}) = \sum_{s=-\infty}^{\infty} \hat{R}_i(s)\lambda_{Ab^{1/3}}(s)e^{-jw_0 s}, \tag{45}$$

where $\hat{R}_i(s) = b^{-1}\sum_{k=i}^{i+b-|s|} X_k X_{k+|s|}$, if $0 \leq |s| \leq b$, and $\hat{R}_i(s) = 0$, if $|s| > b$.

In other words, to calculate $T(\mathbf{X}^{\star(i)})$ we focus attention on the size b subsample $\mathbf{X}^{\star(i)}$, and all data belonging to other subsamples are ignored. Thus, $\hat{R}_i(s)$ is the estimated autocovariance at lag s, where only observations in the subsample $\mathbf{X}^{\star(i)}$ are used in the estimation; similarly, since we chose $M = AN^{1/3}$ as the cut-off parameter in Bartlett's kernel when the sample size was N, we chose $Ab^{1/3}$ as the cut-off parameter when considering our subsamples of size b.

Using (45) we can calculate $T(\mathbf{X}^{\star(i)})$ for $i = 1, ..., B$ (with $B = N - b + 1$), and employ equations (30), (31), (32), and (33) to get accurate estimators of bias, variance, and distribution if the sample size N is large enough. Consequently, (35) would give valid $(1-\alpha)100\%$ confidence intervals for $\theta = f(w_0)$; also, because of the asymptotic normality of $T(\mathbf{X})$, the intervals (34) would also have the correct $(1-\alpha)100\%$ coverage of the unknown $\theta = f(w_0)$ asymptotically. Note that, using the subsampling methodology just described, a $(1-\alpha)100\%$ uniform confidence band for the *whole* unknown function $f(w)$, $w \in [0, 2\pi]$ can be constructed with almost no extra effort; see [46]. Having such a confidence band would, for instance, immediately permit us to test the hypothesis that the spectral density is of a conjectured shape, i.e., to test whether $f(w) = f_0(w)$, for all $w \in [0, 2\pi]$, where $f_0(.)$ is a function of interest; for example, taking $f_0(.)$ to be a constant function leads to a test for 'whiteness' of the X-sequence. The test would reject the hypothesis $f = f_0$ if it were observed that $f_0(w)$ is not covered (for all $w \in [0, 2\pi]$) by the constructed uniform confidence band.

3.7 Some concluding remarks. Although confidence intervals for the spectral density example discussed in Subsection 3.6 can be constructed using various methods, the beauty of the resampling and subsampling data analysis methodology is its simplicity and generality. Thus, the moral of the 'bootstrap philosophy' as described so far can be summarized as follows: *If the general statistic T can be computed from the sample $\mathbf{X}$, then it can certainly be re-computed from pseudo-samples (subsamples, resamples, etc.) and this is enough to gain valuable information on the accuracy of $T(\mathbf{X})$ as a point estimate of θ –by an implicit estimation of the variability*

of $T(\mathbf{X})$ *across samples.* In addition, it can be argued that the bootstrap and jackknife are most useful in nonparametric situations where a parametric model is not available to guide our calculations, and we have to let "the data do all the talking".

In terms of comparing resampling to subsampling, what could be said in general is that subsampling is more widely applicable, covering even situations where the bootstrap fails. The realm of applicability of resampling is nevertheless quite vast, and it can be said that: "when the bootstrap works, it works very well indeed", outperforming other concurrently available methods such as the normal approximation. Looking in particular at the i.i.d. case (elaborated upon in our Sections 1 and 2) where we have i.i.d. data $X_1, X_2, ..., X_N$, and if the statistic of interest $T(\mathbf{X})$ is linear, then the bootstrap will work only if $T(\mathbf{X})$ is asymptotically normal with variance asymptotically proportional to $1/N$; if not, then subsampling *must* be used instead. The same general 'rule of thumb' can be applied in the case where we have weakly dependent stationary observations $X_1, X_2, ..., X_N$, and the statistic of interest is the sample mean (or similarly well-behaved statistics – see, e.g., [40] for an extension of the notion of a linear statistic calculated from dependent data): the (moving blocks) bootstrap will work only if the large sample distribution of our statistic is normal; if not, then subsampling (in its blockwise form of Subsection 3.4) must be used instead.

4. Some bibliographical comments

As of this moment, there are six published books on the bootstrap: the original monograph of Efron [16]; the textbook by Hall [25] that contains a lot of material concerning the higher order accuracy of the bootstrap and the effects of 'studentization'; the collection of research papers in LePage and Billard [33] that also contains an introduction to bootstrap ideas by Efron and LePage; the textbook by Efron and Tibshirani [20] which presents the bootstrap methodology and its applicability in complex data analysis problems – this is definitely a book that the interested reader should consult at some point; the book by Hjorth [27]; and the more recent textbook by Shao and Tu [51] which succeeds in wrapping up all recent theoretical results that are related to the bootstrap and the jackknife. There are also three collections of lecture notes: Beran and Ducharme [5] provide theoretical expositions of the concept of 'prepivoting', a method related to 'studentization', and of bootstrap balanced confidence intervals and prediction regions; Mammen [36] focuses mainly on the bootstrap for linear models, and contains details on a bootstrap variation, the 'wild' bootstrap, which was introduced in [56] – see also [4] in that regard; and Barbe and Bertail [2] consider -among other things- the 'weighted bootstrap' which is an interesting

generalization of the standard bootstrap.

Several review articles are also available in the literature: Efron and Gong [18] and Efron and Tibshirani [19] have a more applied flavor, whereas DiCiccio and Romano [13] give a theoretical treatment; see also the interesting reviews by Efron [15] [17], and Hinkley [26]. Swanepoel [53], and Léger, Politis, and Romano [31] review more recent developments and provide discussion on more advanced applications of the bootstrap methodology; both papers also contain an extensive list of references. Carlstein [12], Léger *et al.* [31], Bose and Politis [11], and Li and Maddala [34] provide reviews of the bootstrap for time series data, while the case of spatial data and random fields is addressed in the research articles by Politis and Romano [40], [43], [44], and Sherman and Carlstein [52]. The reference for most of our Section on subsampling is Politis and Romano [43] that also contains a good number of interesting examples where the bootstrap does *not* work; a critical account of bootstrap ideas was recently presented in [57]. Some specific signal processing applications of the bootstrap are presented in Thomson and Chave [54] and Politis *et al.* [45]. However, for a detailed overview of the bootstrap and its use in signal processing, the review paper by Zoubir and Boashash [58] in this issue of SP Magazine is offered.

References

[1] Athreya, K. (1987), Bootstrap of the mean in the infinite variance case, *Ann. Statist.*, vol 15, pp. 724-731.

[2] Barbe, Ph. and Bertail, P. (1995), *The Weighted Bootstrap*, Lecture Notes in Statistics # 98, Springer Verlag, New York.

[3] Beran, R. (1984). Bootstrap methods in statistics. *Jber. d. Dt. Math.-Verein* **86**, 14-30.

[4] Beran, R. (1986). Discussion to Wu, C.F.J.: Jackknife, Bootstrap and other resampling methods in regression analysis, *Ann. Statist.*, vol. 14, 1295-1298.

[5] Beran, R. and Ducharme, G.R. (1991), *Asymptotic Theory for Bootstrap Methods in Statistics*, Les Publications CRM, Montreal.

[6] Bertail, P. (1994), Second order properties of an extrapolated bootstrap without replacement: the i.i.d. and the strong mixing cases, CORELA-INRA paper no. 9407, (also to appear in *Bernoulli*).

[7] Bertail, P. , Politis, D.N., and Romano, J.P. (1995), On subsampling estimators with unknown rate of convergence, Technical Report No. 95-20, Dept. of Statistics, Purdue University, (also to appear in *J. Amer. Statist. Assoc.*).

[8] Bickel, P. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196-1217.

[9] Bickel, P., Götze, F. and van Zwet, W.R. (1994). Resampling fewer than n observations: Gains, losses and remedies for losses. Technical Report n^0 419, University of California, Berkeley.

[10] Bose, A. (1988), Edgeworth correction by bootstrap in autoregressions, *Ann.Statist.*, 16, pp. 1709-1722.

[11] Bose, A. and Politis, D.N. (1996), A review of the bootstrap for dependent samples, in *Stochastic Processes and Statistical Inference*, B.R. Bhat and B. L. S. Prakasa Rao (Eds.), New Age International Publishers, New Delhi, 1995, pp. 39-51.

[12] Carlstein, E. (1992), Resampling techniques for stationary time series: some recent developments, in *New Direction in Time Series Analysis*, Brillinger, D.R. *et al.* (eds.), Springer Verlag, New York.

[13] DiCiccio, T., and Romano, J. (1988), A review of bootstrap confidence intervals (with discussion), *J. Roy. Statist. Soc., Ser. B*, vol. 50, 338-370.

[14] Efron, B. (1979a), Bootstrap Methods: Another Look at the Jackknife, *Ann. Statist.*, 7, 1-26.

[15] Efron, B. (1979b), Computers and the theory of statistics: thinking the unthinkable, *SIAM Review*, vol. 21, no. 4, pp. 460-480.

[16] Efron, B. (1982),*The Jackknife, the Bootstrap, and other Resampling Plans*, SIAM NSF-CBMS, Monograph 38.

[17] Efron, B. (1988), Computer-intensive methods in statistical regression, *SIAM Review*, vol. 30, no. 3, pp. 421-449.

[18] Efron, B., and Gong, G. (1983), A leisurely look at the Bootstrap, the Jackknife, and Cross-Validation, *Amer. Statistician*, vol. 37, No. 1, pp. 36-48.

[19] Efron, B. and Tibshirani, R.J. (1986), Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy, *Statist. Sci.* **1**, 54-77.

[20] Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[21] Franke, J. and Härdle , W. (1992), On bootstrapping kernel spectral estimates, *Ann.Statist.*, 20, 121-145.

[22] Giné, E. and Zinn, J. (1989), Necessary conditions for the bootstrap of the mean, *Ann.Statist.*, vol 17, pp. 684-691.

[23] Götze, F. and Künsch, H.R.(1993), Second order correctness of the blockwise bootstrap for stationary observations, Preprint 93-061, SFB 343, Bielefeld University.

[24] Hall, P. (1988), Theoretical Comparison of Bootstrap Confidence Intervals, *Ann. Statist.*, 16, 927-953.

[25] Hall, P.(1992), *The Bootstrap and Edgeworth Expansion*, Springer Verlag, New York.

[26] Hinkley, D.V. (1988), Bootstrap methods (with discussion), *J. Roy. Statist. Soc., Ser. B*, vol. 50, 321-337.

[27] Hjorth, J.S.U. (1994), *Computer intensive statistical methods: validation model selection and bootstrap*, Chapman and Hall, New York.

[28] Kreiss, J.P., and Franke, J. (1992), Bootstrapping Stationary Autoregressive Moving Average Models, *J. Time Ser. Anal.*, 13, pp. 297-319.

[29] Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist* **17**, 1217-1241.

[30] Lahiri, S.N.(1992), Edgeworth correction by 'moving block' bootstrap for stationary and nonstationary data, In *Exploring the limits of bootstrap*, ed. by LePage and Billard, John Wiley, pp. 183-214,

[31] Léger, C., Politis, D.N., and Romano, J.P. (1992), Bootstrap Technology and Applications, *Technometrics*, vol. 34, pp. 378-399 .

[32] Lehmann, E.L. (1983), *Theory of point estimation*, John Wiley.

[33] LePage, R. and Billard, L. (eds.) (1992), *Exploring the Limits of Bootstrap*, John Wiley.

[34] Li, H. and Maddala, G.S., (1996), Bootstrapping Time Series Models, *Econometric Rev.*, vol. 15, No. 2, 1996, pp. 115-158.

[35] Liu, R. Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the limits of bootstrap*, ed. by LePage and Billard, John Wiley, pp. 225-248..

[36] Mammen, E. (1992), *When does bootstrap work? asymptotic results and simulations*, Lecture notes in Statistics # 77, Springer, New York.

[37] Miller, R. (1986), *Beyond ANOVA: Basics of Applied Statistics*, John Wiley.

[38] Paparoditis, E. (1996), Bootstrapping Autoregressive and Moving Average Parameter Estimates of Infinite Order Vector Autoregressive Processes, *J. Multivar. Anal.*, vol. 57, pp. 277-296.

[39] Politis, D.N., and Romano, J.P. (1992). A circular block-resampling scheme for stationary data, In *Exploring the limits of bootstrap*, ed. by LePage and Billard, John Wiley, pp. 263-270.

[40] Politis, D.N., and Romano, J.P. (1993a). Nonparametric Resampling for Homogeneous Strong Mixing Random Fields, *J. Multivar. Anal.*, vol. 47, No. 2, pp. 301-328.

[41] Politis, D.N., and Romano, J.P. (1993b). 'Estimating the Distribution of a Studentized Statistic by Subsampling', *Bulletin of the International Statistical Institute*, 49th Session, Firenze, August 25 - September 2, 1993, Book 2, pp. 315-316.

[42] Politis, D.N., and Romano, J.P. (1994a), The Stationary Bootstrap, *J. Amer. Statist. Assoc.*, vol. 89, No. 428, pp. 1303-1313.

[43] Politis, D.N., and Romano, J.P. (1994b), Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions, *Ann.Statist.*, vol. 22, No. 4, 2031-2050.

[44] Politis, D.N., and Romano, J.P. (1996), 'Subsampling for Econometric Models – Comments on Bootstrapping Time Series Models', *Econometric Rev.*, vol. 15, No. 2, 1996, pp. 169-176.

[45] Politis, D.N., Romano, J.P., and Lai, T.-L. (1992), 'Bootstrap Confidence Bands for Spectra and Cross-Spectra', *IEEE Trans. Signal Proc.*, vol. 40, No. 5, May 1992, pp. 1206-1215.

[46] Politis, D.N., Romano, J.P. and You, L. (1993), 'Uniform confidence bands for the spectrum based on subsamples', in *Computing Science and Statistics, Proceedings of the 25th Symposium on the Interface*, San Diego, California, April 14-17, 1993, (M. Tarter and M. Lock, eds.), The Interface Foundation of North America, pp. 346-351.

[47] Priestley, M. B. (1981), *Spectral Analysis and Time Series*, Academic Press.

[48] Quenouille, M. (1949), Approximate tests of correlation in time series, *J. Roy. Statist. Soc. (Ser. B)*, 11, pp. 68-84.

[49] Radulovic, D. (1996), The bootstrap of the mean for strong mixing sequences under minimal conditions, *Statist. Prob. Letters,* 28, pp. 65-72.

[50] Singh, K.(1981), On the asymptotic accuracy of Efron's bootstrap, *Ann.Statist.*, 9, 1187-1195.

[51] Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer, New York.

[52] Sherman, M. and Carlstein, E. (1996), Replicate histograms, *J. Amer. Statist. Assoc.*, 91, No. 434, 566-576.

[53] Swanepoel, J.W.H. (1990), A review of bootstrap methods, *South African Statist. J.*, vol. 24, pp. 1-34.

[54] Thomson, J. and Chave, A.D. (1991), Jackknifed error estimates for spectra, coherences, and transfer functions, in *Advances in Spectrum Analysis and Array Processing, vol. I*, S. Haykin (ed.), Prentice-Hall, New Jersey.

[55] Tukey, J.W. (1958), Bias and confidence in not quite large samples, (abstract), *Ann. Math. Statist.*, 29, p. 614.

[56] Wu, C.F.J. (1986) Jackknife, Bootstrap and other resampling methods in regression analysis, *Ann. Statist.*, vol. 14, 1261-1295.

[57] Young, G.A. (1994), Bootstrap: more than a stab in the dark? (with discussion), *Statist. Sci.* **9**, No. 33, 382-415.

[58] Zoubir, A.M. and Boashash, B. (1997), The bootstrap: theory and signal processing applications, *Signal Proc. Magazine*, same issue.