A Primer on Bootstrap Methods in Statistics

by

Dimitris N. Politis
Purdue University

Technical Report #95-19

Department of Statistics
Purdue University

April 1995

*

A primer on bootstrap methods in statistics *

Dimitris N. Politis

Department of Statistics
Purdue University
W. Lafayette, IN 47907

Abstract

A tutorial introduction to the recently developed resampling methods for statistical inference, i.e., the bootstrap, the jackknife, and some of their variations, is presented. The objective is to establish the basic ideas of the methodology so that the interested reader can proceed to refer to the review papers, books, and monographs listed in the bibliography.

1. Resampling and the bootstrap

The goal of this article is to provide a readable, self-contained introduction to the bootstrap and jackknife methodology for statistical inference intended for people that have not been previously exposed to the subject; in particular, the focus is on the derivation of confidence intervals in general situations.

1.1 The general non-parametric set-up. Suppose that $X = (X_1, \ldots, X_N)$ 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

*This is an extensive revision of Purdue Technical Report #93-49.

sample is studied in order to estimate a certain parameter $\theta(F)$ associated with the distribution F whose form is unknown. A statistic $T = T(X)$ might be used to estimate $\theta(F)$ from the data. However, *a measure of the statistical accuracy of the point estimator $T(X)$ 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 $T(X)$ from $\theta(F)$, i.e., the 'error' in estimating $\theta(F)$ by $T(X)$, could be statistically quantified; in that case, the practitioner would be able to gauge how much importance to attach to the individual 'measurement' $T(X)$. For example, the bias (='systematic error') and the variance (='random error') of the estimator T are of interest, and are defined as follows:

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

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

where E_F denotes expectation under the F distribution. The quantity $T(X) - \theta(F)$, i.e., the estimator minus the estimand, represents the 'error' in estimating $\theta(F)$ by $T(X)$; it is also sometimes called a 'root', since if the estimator is accurate, $T(X) - \theta(F)$ should be close to zero (cf., for example, DiCiccio and Romano (1988), or Beran and Ducharme (1991)). 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 that $\theta(F)$ is a location parameter, say the mean or median of F, and $T(X)$ 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 $T(X)$ 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 $T(X)$.

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

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

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

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

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

Estimates of $Bias_F(T)$ and $Var_F(T)$ might be available in the statistical literature for different problems. For example, if $T(X) = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ 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 $T(X)$ 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. Lehmann (1983)), but are substantially more complicated.

The bootstrap (and the closely related jackknife (cf. Efron (1979, 1982)) could alternatively be used to easily obtain estimates of $Bias_F(T)$ and $Var_F(T)$ for a wide variety of statistics $T(X)$. However, before going into that, let us look at this problem from a different angle.

1.3 The usefulness of Monte Carlo randomization. Suppose, for the sake of argument, that the population and its distribution F were in fact known. Then, $Bias_F(T)$ and $Var_F(T)$ could be calculated exactly by analytical methods, or approximately by Monte Carlo simulation, in case the analytical computation is difficult.

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

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, $X^{(1)}, \ldots, X^{(B)}$, 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

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

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

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

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

But if the population is considered known, we could also directly evaluate the sampling distribution of the root $T(X) - \theta(F)$, 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

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

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

$$[T(X) - q(1-\alpha/2), T(X) - q(\alpha/2)], \tag{9}$$

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. 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. Hall (1988, 1992)) and possess some interesting theoretical properties; nevertheless, the confidence intervals that are most often used in practice are equal-tailed (cf. Efron and Tibshirani (1993)).

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

$$Dist_{T,F}(x) = P_F(T(X) \leq x) \tag{10}$$

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

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

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

$$Dist_{T,F}(x) \simeq \frac{1}{B} \sum_{i=1}^{B} 1(T(X^{(i)}) \leq x) = \frac{1}{B}(\#T(X^{(i)}) \leq x) \tag{11}$$

i.e., the theoretical probability should be approximately equal to the observed sample proportion if B is large.[2] From equation (11), the quantiles of $Dist_{T,F}(x)$ (and therefore also those of $Dist_{T,F,\theta}(x)$)) can be approximately calculated, and the confidence interval (9) constructed.

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 $Dist_{T,F,\theta}(x)$. 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* $X = (X_1, \ldots, X_N)$.

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 (1979). 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 $X^{*(1)}, \ldots, X^{*(B)}$ (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.

[2] Note that $(\#T(X^{(i)}) \leq x)$ reads: number of the $T(X^{(i)})$'s among $T(X^{(1)}), \ldots, T(X^{(B)})$ that are observed to be less or equal to x; equation (11) should be viewed as describing a function of the real argument x, and can be plotted as such.

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

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

where x is any real number. To elaborate, in order to form the ith resample $X^{*(i)} = (X_1^{*(i)}, \ldots, X_N^{*(i)})$, 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 (cf. equations (1), (2)) by

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

$$Var^*(T) = Var_{\hat{F}}(T) = E_{\hat{F}}T^2(X) - [E_{\hat{F}}T(X)]^2. \tag{14}$$

It should be noted that $\theta(\hat{F})$ is just the sample statistic corresponding to the population parameter $\theta(F)$. In most cases, the statistic $T(X)$ is chosen to be 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. $T(X) = \theta(\hat{F})$. Unless otherwise stated, we will henceforth assume that $\theta(\hat{F}) \equiv T(X)$ for simplicity; in a different situation the 'plug-in' principle can be appropriately modified.

1.6 A parametric set-up. This 'plug-in' viewpoint permits one to see how the bootstrap would work in a parametric problem as well. For example, if the distribution $F(x) = F_\theta(x)$ is known up to some parameter θ, then 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, where $\hat{\theta} = T(X)$ is the estimated (from our sample) value of the parameter θ. All the Monte Carlo approximations remain valid, except that in the parametric set-up, to form the ith resample $X^{*(i)} = (X_1^{*(i)}, \ldots, X_N^{*(i)})$, 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. Miller (1986)); however, the bootstrap will tend to give more accurate estimates in general (cf. Hall (1992), Efron and Tibshirani (1993)). Having said that, let us return and focus our attention on the general non-parametric 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:

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

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

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 $T(X) - \theta(F)$ given in equation (8).

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

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

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

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

where $q^*(\alpha/2)$ and $q^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{T,\hat{F},\theta}(x)$ distribution respectively. It should be noted 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 Efron and Tibshirani (1993, chapter 22) for a thorough discussion, and DiCiccio and Romano (1988), Hall (1988, 1992) for a comparison of bootstrap confidence intervals. Note that in the terminology of Hall (1988), equation (18) represents a confidence interval based on the 'hybrid' method, whereas in Hall (1992) equation (18) is the 'other percentile method' confidence interval. We will also refer to equation (18) as the *hybrid* method for bootstrap confidence intervals; the hybrid method is a cross between the percentile and the bootstrap-t methods that will be defined shortly.

The bootstrap distributions $Dist_{T,\hat{F},\theta}(x)$ and $Dist_{T,\hat{F}}(x) = P_{\hat{F}}(T(X) \leq x)$ –and therefore their quantiles as well– can be easily evaluated by Monte Carlo as follows:

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

and

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

Similarly to equation (11), the functions $Dist_{T,\hat{F},\theta}(x)$ 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 $T(X^{*(i)})$, for $i = 1, ..., B$; this histogram would be an approximation to the probability *density* function of the random variable $T(X)$, while $Dist_{T,\hat{F}}(x) = P_{\hat{F}}(T(X) \leq x)$ is an approximation to the *cumulative* probability distribution function of $T(X)$. The article by Zoubir and Boashash (1995) contains many examples of such plotted histograms that are very helpful in terms of visual inspection and interpretation of the variability of $T(X)$.

Note that the probability distribution function $Dist_{T,\hat{F}}(x)$ is described by the simple equation (19), and actually represents the 'empirical' distribution function of the observed $T(X^{*(i)})$, or, in other words, the distribution function of the (pseudo) sample consisting of $T(X^{*(i)})$, $i = 1, ..., B$. The graph of $Dist_{T,\hat{F}}(x)$ 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 $T(X^{*(i)})$. These observations point to a very easy way of figuring out the quantiles of distribution $Dist_{T,\hat{F}}(x)$ from which the quantiles of $Dist_{T,\hat{F},\theta}(x)$ can be readily obtained; for, if $q^*(\alpha/2)$ and $q^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{T,\hat{F},\theta}(x)$ distribution,

then

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

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 the $Q^*(\alpha/2)$ and $Q^*(1-\alpha/2)$ quantiles. Start by sorting the values $T(X^{*(i)})$, $i = 1, ..., B$, 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

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

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

$$Q^*(\alpha/2) = T_{k_1}^*, \quad Q^*(1-\alpha/2) = T_{k_2}^*, \tag{22}$$

where $k_1 = [B\alpha/2] + 1$, $k_2 = [B(1-\alpha/2)] + 1$, and $[.]$ denotes the integer part.

Putting equations (21) and (22) together, we obtain a quick-and-easy alternative formulation of the hybrid confidence interval of equation (18) as

$$[2T(X) - T_{k_2}^*, 2T(X) - T_{k_1}^*]. \tag{23}$$

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

$$[T_{k_1}^*, T_{k_2}^*]. \tag{24}$$

The percentile bootstrap confidence intervals are also very popular (as popular as the hybrid intervals, if not more); however, the justification for their use is not obvious from what has been discussed so far here. Rather than going into more details, we can refer the reader to the very interesting exposition in Efron and Tibshirani (1993, p. 170 and on); also look at Example 1 and Table 2 of Zoubir and Boashash (1995) where the percentile intervals are employed.

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 $\hat{F}$, 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 Zoubir and Boashash (1995) 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, 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[3] estimator of the variance $Var_F(T)$; let us call this estimator $\widehat{Var}_F(T)$. To fix ideas, consider the simplest case where the statistic $T(X)$ of interest is the sample mean $\bar{X}$. In this case, 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 $T(X)$ by its estimated standard deviation $\sqrt{\widehat{Var}_F(T)}$ is usually referred to as 'studentization', since –*if* the data were Gaussian– this would result in Student's t-distribution. Consider then the sampling distribution of the 'studentized' statistic, i.e.,

$$Dist_{student}(x) = P_F\left(\frac{T(X) - \theta(F)}{\sqrt{\widehat{Var}_F(T)}} \leq x\right). \tag{25}$$

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

$$[T(X) - u(1-\alpha/2)\sqrt{\widehat{Var}_F(T)}, T(X) - u(\alpha/2)\sqrt{\widehat{Var}_F(T)}], \tag{26}$$

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

[3]Loosely speaking, an estimator is consistent if it is accurate for large samples, i.e., *asymptotically* correct.

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 $Dist_{student}(x)$ and its quantiles are unknown; nevertheless, it can be estimated by the bootstrap, similarly to what was discussed in the previous sections. In particular, the bootstrap distribution $Dist^*_{student}(x)$ that can be used to approximate $Dist_{student}(x)$ is given by

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

$$= \frac{1}{B}(\#T(X^{*(i)}) \leq x\sqrt{\widehat{Var}_F{}^{*(i)}(T)} + \theta(\hat{F})), \tag{27}$$

where $\widehat{Var}_F{}^{*(i)}(T)$ is the estimate of the variance of the statistic $T(X)$ *as computed from the* $X^{*(i)}$ *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 section 1.3; in that case, $\widehat{Var}_F{}^{*(i)}(T)$ is the bootstrap variance estimate computed from the $X^{*(i)}$ resample! In other words, we have an *iterated* or *nested* bootstrap –a bootstrap simulation for each of the original bootstrap resamples; cf. Hall (1992) or Efron and Tibshirani (1993) for more details.

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

$$[T(X) - u^*(1-\alpha/2)\sqrt{\widehat{Var}_F(T)}, T(X) - u^*(\alpha/2)\sqrt{\widehat{Var}_F(T)}], \tag{28}$$

where $u^*(\alpha/2)$ and $u^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap $Dist^*_{student}(x)$ distribution respectively; the confidence interval of equation (28) is called a bootstrap-t or a percentile-t interval due to the 'studentization'.

As it turns out (cf. Singh (1981)), the confidence interval of equation (28) is *more* accurate than *either* the hybrid 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', or that the bootstrap 'captures the skewness' of the underlying distribution. The mathematical explanation of this phenomenon is based on the theory of Edgeworth expansions of $Dist_{student}(x)$ and $Dist^*_{student}(x)$ in powers of $1/\sqrt{N}$, and can be found in Hall (1992); see also Example 5 and the Appendix of Zoubir and Boashash (1995) where 'studentization' (also called 'pivoting') is

further discussed. 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), (28), etc. have coverage probability approximately equal to $(1-\alpha)100\%$ as they are supposed to. Although we do not attempt to give a definite 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. In other words, it seems that the bootstrap would *not* work unless $T(X)$ is indeed asymptotically normal[4]. However, asymptotic normality of $T(X)$ is not a sufficient condition for the validity of the bootstrap; other assumptions on the particular problem at hand must also hold, e.g., moment assumptions. For example, it seems to be the case that the bootstrap would *not* work unless the variance of $T(X)$ is (for large N) approximately proportional to $1/N$; more on this property may be found in our Section 2.1.

1.9 Trasformations and variance stabilization. The reader should also refer to the textbook by Efron and Tibshirani (1993) 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 hybrid 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

[4]This point of view allows for a heuristic explanation regarding how the hybrid confidence interval of equation (23), and the percentile interval of equation (24) could be (and are) simultaneously valid. The reason is that, if the root $T(X)-\theta(\hat{F})$ is approximately normal $N(0, Var_F(T))$, then $\theta(\hat{F})-T(X)$ 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 $T(X)-\theta(\hat{F})$ and of $\theta(\hat{F})-T(X)$ are the *same*; thus the confidence intervals (23) and (24) are both correct (at least to first order).

by the percentile intervals of (24), but *not* shared by either the hybrid intervals of (18), or the bootstrap-t intervals of (28), 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 $T = T(X)$ 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), hybrid bootstrap of equation (18), percentile bootstrap of equation (24), bootstrap-t of equation (28), 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 better?", 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. Efron and Tibshirani (1993, p. 54 and p. 163)), a transformation is available in the literature that approximately 'normalizes' and 'variance stabilizes' the estimator $T(X)$; in other words, the estimator $g(T)$ has a distribution that is closer to being Gaussian than the distribution of $T(X)$, 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 $T(X)$ 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. Miller (1986),

Efron and Tibshirani (1993)). 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 section on studentization– to calculate estimates of $Var_F(T)$ from each resample; details can be found in Efron and Tibshirani (1993, p. 163), and in Zoubir and Boashash (1995). 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 hybrid 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 Section 1.8, in the case where the statistic $T(X)$ 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 (1979) pioneering paper, with main focus to show the applicability of the bootstrap in many different settings; Bickel and Freedman (1981) and Beran (1984) are two very important early papers in this connection (see also our Section 4 for more bibliographical comments).

At another level, recall that performing the bootstrap in practice requires sampling *with* replacement from the observations $X_1, \ldots, X_N$, to get a resample of size N. The *exact* computation of the bootstrap distribution would actually involve taking into account *all* the possible resamples, weighted by the corresponding multinomial probabilities; however, the number of possible resamples is $\frac{(2N-1)!}{N!(N-1)!}$ which is impractically large. Doing the Monte Carlo random bootstrap sampling gets around this problem, but there is also another way of lowering the computational complexity: the jackknife and subsampling.

2.1 The jackknife idea. Consider sampling *without* replacement from the observations $X_1, \ldots, X_N$, to get a resample (now called a *subsample*) of size b, where of course $b < N$. If $b = N - 1$, this is exactly the original jackknife of Quenouille and Tukey (cf. Efron (1979, 1982) and Efron and Tibshirani (1993) for details), and there are only N possible different subsamples. 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 $T(X)$; these will be given in a more general form in what follows.

In general, one can take an arbitrary b, not necessarily equal to $N-1$, yielding the so-called delete-d jackknife, where $d = N - b$; see Shao and Tu (1995) for more details. Observe that the number of possible subsamples now rises to $\frac{N!}{b!(N-b)!}$ 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. As long as b is large enough (but of smaller order of magnitude than N) the subsampling distribution estimates are asymptotically correct; see Politis and Romano (1995).

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, even in cases where the statistic $T(X)$ 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 $T(X)$ is approximately c^2/τ_N^2, for large N, where c is some constant, and τ_N is a function of the sample size N; in regular cases, e.g., if $T(X)$ is the sample mean, sample median, sample variance, etc., we have that $\tau_N = \sqrt{N}$. 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:

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

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

$$Bias^\star(T) \simeq \frac{\tau_r}{\tau_N}\left(\frac{1}{B}\sum_{i=1}^{B} T(X^{\star(i)}) - T(X)\right) \tag{29}$$

$$Var^\star(T) \simeq \frac{\tau_r^2}{\tau_N^2}\left(\frac{1}{B}\sum_{i=1}^{B} T^2(X^{\star(i)}) - [\frac{1}{B}\sum_{i=1}^{B} T(X^{\star(i)})]^2\right) \tag{30}$$

$$Dist^\star_{T,F}(x) \simeq \frac{1}{B}\sum_{i=1}^{B} 1(T(X^{\star(i)}) \le x\frac{\tau_N}{\tau_r}) = \frac{1}{B}(\#T(X^{\star(i)}) \le x\frac{\tau_N}{\tau_r}) \tag{31}$$

and

$$Dist^\star_{T,F,\theta}(x) \simeq \frac{1}{B}(\#T(X^{\star(i)}) \le x\frac{\tau_N}{\tau_r} + T(X)), \tag{32}$$

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 τ_b in (29), (30), (31), and (32) 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 $X^{\star(1)},\ldots,X^{\star(B)}$ 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 if 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 (29), (30), (31), and (32) are accurate provided the sample size

N is large, and that one of the following three conditions is met:

(a) The estimator $T(X)$ is of very simple form, e.g., $T(X)$ is the sample mean, or maybe a trimmed mean (but not the sample median!), *and* the population distribution F is *known* to be normal (with some unknown mean and variance); here the ordinary jackknife (with $b = N - 1$) would work (cf. Efron and Tibshirani (1993)).

(b) The estimator is not necessarily that simple, but it satisfies the 'regularity' condition that $\tau_N = \sqrt{N}$; for example, $T(X)$ 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., $b = [kN]$, where k is a constant in $(0, 1)$, and $[.]$ denotes the integer part; see Shao and Tu (1995).

(c) The estimator is arbitrarily complex, 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. $b = [N^k]$, 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 (29), (30), (31), and (32) will be valid with τ_b used instead of τ_r, which is perhaps more intuitive; see Politis and Romano (1995) for more details on this general case.

Under one of conditions (a), (b) or (c) above, the estimates proposed in equations (29), (30), (31), and (32) 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 $T(X)$ is known to be asymptotically normal (see our Section 1.1), then a $(1 - \alpha)100\%$ confidence interval for $\theta(F)$ can be given by

$$[T(X) - Bias^\star(T) - z\sqrt{Var^\star(T)}, T(X) - Bias^\star(T) + z\sqrt{Var^\star(T)}] \tag{33}$$

where $z = z(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of the standard normal distribution as in (4). If $T(X)$ 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

$$[T(X) - q^\star(1 - \alpha/2), T(X) - q^\star(\alpha/2)], \tag{34}$$

where $q^\star(\alpha/2)$ and $q^\star(1 - \alpha/2)$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the $Dist^\star_{T,F,\theta}(x)$ dis-

tribution respectively. Note that if both confidence intervals (33) and (34) are valid, i.e., if $T(X)$ is asymptotically normal, the interval (33) would be considered to be the one that is more accurate; that is, the coverage probability of (33) would be closer to the desired value of $(1-\alpha)100\%$. 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 (34) in particular) is more widely applicable in the setting of condition (c) above; see Section 3.6 for an interesting such example.

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

What has been discussed so far hinges on the assumption that the data $X = (X_1, \ldots, X_N)$ 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,[5] 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.

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 residuals $\hat{\epsilon}_i = e_i - N^{-1}\sum_{k=1}^{N} e_k$[6]. 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

$$\hat{G}(x) = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{\epsilon}_i \leq x) = \frac{1}{N}(\#\hat{\epsilon}_i \leq x).$$

The bootstrap resampling in this case can be done as follows:

[5] 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.

[6] 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.

- Draw B i.i.d. samples $E^{*(1)}, \ldots, E^{*(B)}$ (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 $E^{*(k)} = (\hat{\epsilon}_1^{*(k)}, \ldots, \hat{\epsilon}_N^{*(k)})$, 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 $E^{*(k)}$ 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.,

$$Bias^*(\hat{\beta}) \simeq \frac{1}{B} \sum_{i=1}^{B} \hat{\beta}^{*(k)} - \hat{\beta} \tag{35}$$

$$Var^*(\hat{\beta}) \simeq \frac{1}{B} \sum_{i=1}^{B} (\hat{\beta}^{*(k)})^2 - [\frac{1}{B} \sum_{i=1}^{B} \hat{\beta}^{*(k)}]^2 \tag{36}$$

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

$$[\hat{\beta} - g^*(1 - \alpha/2), \hat{\beta} - g^*(\alpha/2)]; \tag{37}$$

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

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

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) given by

$$X_i = \beta X_{i-1} + \epsilon_i, \tag{38}$$

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[7] is included in the right-hand-side of the auto-regression (38). It is also usually assumed that $|\beta| < 1$, in which case we may consider that the X_i's were obtained recursively by letting $i = 1, ..., N$ in equation (38), and some proper choice of X_0. Note that the data $X_1, ..., X_N$ can be explained by means of the model (38) driven by i.i.d. errors. It is apparent that since the parameters in the model (in this case it is just β) can be estimated 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 Section 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 Section 3.1 applies *verbatim* here as well, although the Y_i are no longer nonrandom; for more details on

[7]Note however that, although the theoretical mean satisfies $EX_i = 0$, 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 also Zoubir and Boashash (1995)).

the bootstrap for linear or nonlinear autoregressive time series models see Léger *et al.* (1992) and Zoubir and Boashash (1995).

3.3 Data that are not too dependent: weakly dependent observations. Suppose that no plausible model (such as the autoregression of Section 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 $T(X)$, 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 (see, e.g., Politis and Romano (1992, 1994)), 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 $X^{\star(i)}$ would consist of the observations $X_i, ..., X_{i+b-1}$, and i runs from 1 to $N-b+1$. 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 Section 2.1. In other words, the statistic T would be evaluated over each of the $B = N-b+1$ subsamples creating the pseudo-replications $T(X^{\star(1)}), \ldots, T(X^{\star(B)})$.

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 Politis and Romano (1995) for more details. Note, however, that similarly to condition (c) in Section 2.2, here we *have* to choose a b such that b is large, but b/N is small, e.g. $b = [N^k]$, where k is a constant in $(0,1)$; standard choices of b would be $b = [N^{1/2}]$ or $b = [N^{1/3}]$. With such a choice for b, equations (29), (30), (31), and (32) would apply here *verbatim* and they would give accurate estimators of bias, variance, and distribution if the sample size N is large enough. Consequently, (34) would give a valid $(1-\alpha)100\%$ confidence interval for $\theta(F)$, whereas (33) would also give a valid $(1-\alpha)100\%$ confidence interval for $\theta(F)$ provided the estimator $T(X)$ 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 Section 3.4. This is the essense of the 'moving blocks' bootstrap that was introduced by Künsch(1989) and Liu and Singh(1992); see also Politis and Romano (1992, 1994) for closely related proposals for bootstrapping dependent data.

Consider the set $\mathcal{S} = \{X^{\star(1)}, X^{\star(2)}, ..., X^{\star(N-b+1)}\}$, where $X^{\star(i)} = (X_i, ..., X_{i+b-1})$ is the ith subsample defined in Section 3.4; so $\mathcal{S}$ is a set of subsamples. As in Section 3.4, here as well we require that b is large, but b/N is small, e.g. $b = [N^k]$, where k is a constant in $(0,1)$.

Let $K = [N/b]$, 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 $X^{**(1)}, \ldots, X^{**(K)}$; concatenate the observations found in $X^{**(1)}, \ldots, X^{**(K)}$ into a series of $L = bK$ observations denoted by $Y^{**(1)}$. Take another random sample of size K with replacement from the set $\mathcal{S}$, and store it in $Y^{**(2)}$. In the same manner, generate $Y^{**(i)}$, for $i = 3, 4, ..., B$. Now evaluate the statistic T over each of the $Y^{**(i)}$, for $i = 1, 2, 3, ..., B$, bootstrap pseudo-series to get the pseudo-replications $T(Y^{**(1)}), ..., T(Y^{**(B)})$.

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 $Dist_{T,F,\theta}(x)$ are $Bias^{**}(T)$, $Var^{**}(T)$, $Dist^{**}_{T,F}(x)$, and $Dist^{**}_{T,F,\theta}(x)$ respectively which are presented below:

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

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

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

and

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

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

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

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 Section 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 Section 1.8. In other words, if $T(X)$ is asymptotically normal, the 'moving blocks' bootstrap can be applied to an appropriately 'studentized' version of $T(X)$ to yield confidence intervals that are more accurate than the intervals obtained from the normal approximation; see, for example, Lahiri (1992). 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 $T(X)$ is (for large N) approximately proportional to $1/N$, *and* that $T(X)$ is indeed asymptotically normal.

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 $f(w) = \frac{1}{2\pi}\sum_{s=-\infty}^{\infty} R(s)e^{-jsw}$; 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

$$\hat{f}(w) = \frac{1}{2\pi}\sum_{s=-\infty}^{\infty} \hat{R}(s)\lambda_M(s)e^{-jws},$$

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

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

It is well known (cf. Priestley (1981)) that, under some regularity conditions, $\hat{f}(w)$ is asymptotically normal, and that $Var(\hat{f}(w)) \approx \frac{2M}{3N} f^2(w)(1+\eta(w))$, 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

$$T(X) \equiv \hat{f}(w_0) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} \hat{R}(s)\lambda_{[AN^{1/3}]}(s)e^{-jw_0 s}. \tag{44}$$

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 $T(X)$ is (for large N) approximately proportional to $N^{-2/3}$, and $\tau_N = N^{1/3}$.*

Since the variance of $T(X)$ 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 by Politis, Romano, and Lai (1992) in order to handle this 'difficult' example. Nonetheless, the subsampling methodology as described in Section 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 Section 3.4. To elaborate, let the ith subsample $X^{\star(i)}$ consist of the observations $X_i, ..., X_{i+b-1}$; applying the statistic T on the subsample $X^{\star(i)}$ amounts to letting

$$T(X^{\star(i)}) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} \hat{R}_i(s)\lambda_{[Ab^{1/3}]}(s)e^{-jw_0 s}, \tag{45}$$

where $\hat{R}_i(s) = b^{-1}\sum_{k=i}^{i+b-1} X_k X_{k+s}$, if $0 \le s \le b$, $\hat{R}_i(s) = 0$, if $s > b$, and $\hat{R}(s) = \hat{R}(-s)$, if $s < 0$. In other words, to calculate $T(X^{\star(i)})$ we focus attention on the size b subsample $X^{\star(i)}$, 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 $X^{\star(i)}$ 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 our sample (the ith subsample) was of size b.

Using (45) we can calculate $T(X^{*(i)})$ for $i = 1, ..., B$ (with $B = N - b + 1$), and employ equations (29), (30), (31), and (32) to get accurate estimators of bias, variance, and distribution if the sample size N is large enough. Consequently, both (34) and (33) would give valid $(1 - \alpha)100\%$ confidence intervals for $\theta = f(w_0)$ since $T(X)$ is known to be asymptotically normal.

Although confidence intervals for this particular problem can be constructed using various methods, the beauty of the subsampling (and resampling) 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 X, then it can certainly be re-computed from pseudo-samples (subsamples, resamples, etc.) of the type $X^{*(i)}$, $i = 1, ..., B$, and this is enough to gain valuable information on the accuracy of $T(X)$ as a point estimate of θ.*

4. Some bibliographical comments

At this moment, there are five published books on the bootstrap: the original monograph of Efron (1982); the textbook by Hall (1992) 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 (1992) that also contains an introduction to bootstrap ideas by Efron and LePage; the textbook by Efron and Tibshirani (1993) 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; and the book by Hjorth (1994). A book by Shao and Tu (1995) is also in press at the moment; this interesting new book attempts to wrap up all the recent theoretical results that are related to the bootstrap and the jackknife. There are also three collections of lecture notes: Beran and Ducharme (1991) provide theoretical expositions of the concept of 'prepivoting', a method related to 'studentization', and of bootstrap balanced confidence intervals and prediction regions; Mammen (1992) focuses mainly on the bootstrap for linear models, and contains details on a bootstrap variation, the 'wild' bootstrap, which was introduced in Wu (1986) – see also Beran (1986) in that regard; and Barbe and Bertail (1995) 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 (1983) and Efron and Tibshirani (1986) have a more applied flavor, whereas DiCiccio and Romano (1988) give a theoretical treatment; see also Hinkley (1988). Swanepoel (1990), and Léger, Politis, and Romano (1992) review more recent developments and provide discussion on more advanced applications of the bootstrap methodology; both papers also contain an extensive list of references. Léger *et al.* (1992) and Bose and Politis (1995) provide reviews of the bootstrap for dependent samples, while the reference for most of our section on subsampling is Politis and Romano (1995) that also contains a good number of examples where the bootstrap does *not* work; a critical account of bootstrap ideas was recently presented in Young (1994). Finally, the review paper by Zoubir and Boashash (1995) should be consulted, especially in connection with signal processing applications of the bootstrap.

Acknowledgement

The present article was written as a part of a tutorial set of notes for students in the course STAT526 offered by the Department of Statistics of Purdue University. Many thanks are due to Professor Mary Ellen Bock (Purdue University) for her critical reading of the manuscript; the support of NSF grant DMS94-04329 is also gratefully acknowledged.

References

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

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

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

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

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

[6] Bose, A. and Politis, D.N. (1995), A review of the bootstrap for dependent samples, to appear in *Stochastic Processes and Statistical Inference*, B.R. Bhat and B. L. S. Prakasa Rao (Eds.), Wiley Eastern, New Delhi.

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

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

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

[10] 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.

[11] 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.

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

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

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

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

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

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

[18] 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,

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

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

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

[22] 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..

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

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

[25] 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.

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

[27] Politis, D.N., and Romano, J.P. (1995), Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions, to appear in *Ann.Statist.*.

[28] 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.

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

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

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

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

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

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

[35] Zoubir, A.M. and Boashash, B. (1995), The bootstrap: theory and signal processing applications, *Signal Proc. Magazine.*