The jackknife and the bootstrap for general stationary observations

Hans R. Künsch

ETH Zürich

Summary

We extend the jackknife and the bootstrap method of estimating standard errors to the case where the observations form a general stationary sequence. We do not attempt a reduction to i.i.d. values. The jackknife calculates the sample variance of replicates of the statistic obtained by omitting each block of ℓ consecutive data once. In the case of the arithmetic mean this is shown to be equivalent to a weighted covariance estimate of the spectral density of the observations at zero. Under appropriate conditions consistency is obtained if $\ell = \ell(n) \to \infty$ and $\ell(n)/n \to 0$. General statistics are approximated by an arithmetic mean. In regular cases this approximation determines the asymptotic behavior. Bootstrap replicates are constructed by selecting blocks of length ℓ randomly with replacement among the blocks of observations. The procedures are illustrated by using the sunspot numbers and some simulated data.

AMS 1980 subject classification: Primary 62G05, 62G15, Secondary: 62M10.

Key Words and Phrases: Variance estimation, jackknife, bootstrap, statistics defined by functionals, time series, influence function.

1. Introduction

The jackknife (Tukey, 1958) and the bootstrap (Efron, 1979) have become well established as nonparametric estimators of the variance of a statistic. However the assumption of independence of the observations is crucial. It is easily seen that they give incorrect answers if dependence is neglected, compare Remark 2.1 of Singh (1981). Recently the two methods have been extended to ARMA-models by reducing to innovations which are i.i.d., see Davis (1977), Freedman (1984), Efron and Tibshirani (1986), Section 6. Still ARMA-processes are not able to model essential features of many observed time series, compare Priestley (1981), Chapter 11. Fitting models which go beyond ARMA is however an extremely difficult task, and it seems impossible to take the effects of parameter estimation or misspecification of the model into account. Moreover a variance estimator can be unreliable even if the true distribution differs only slightly from the model and the statistic is robust, see Section 2.4 below.

Because of these reasons we propose here an extension of the standard jackknife and bootstrap which does not require us to fit a parametric or semiparametric model first. It works for arbitrary stationary processes with short range dependence expressed for instance with mixing conditions. For the jackknife we delete each block of ℓ consecutive observations once and calculate the sample variance of the values of the statistic obtained in this way. Moreover we make a smooth transition between observations left out and observations with full weight similarly to tapering in time series analysis. For the bootstrap we choose n/ℓ blocks of length ℓ with replacement from the $n - \ell + 1$ blocks of observed data.

If the statistic is not a symmetric function of the observations, leaving out observations in the middle or joining randomly selected blocks causes problems. Our definition in Chapter 2 takes care of this, but we have to restrict ourselves to statistics which are given by a functional of an empirical marginal with fixed dimension. In Chapter 3 we show that these procedures are consistent in the case of the arithmetic mean and obtain the asymptotic bias and variance. In this case the jackknife reduces to a standard spectral estimation procedure. In Chapter 4 we study general statistics by von Mises expansions. We show that for smooth functionals the linear approximation completely determines the asymptotic behavior of the jackknife. Chapter 5 contains examples with real and simulated data.

Carlstein (1986) has proposed a variance estimator which selects non- overlapping blocks. For the arithmetic mean deletion of blocks is the same as selecting blocks. So in this case our jackknife differs only by using overlapping blocks and tapering. However for general statistics, deletion is better than selection, both in theory (see Remark 4.1 below) and in the simulations of Sections 5.l and 5.2.2.

2. Definitions

When we try to formalize the intuitive ideas from the Introduction, some problems occur. For the jackknife we need to define the statistics with a missing block of observations, and with the bootstrap we have to take care how we join two randomly selected blocks. These difficulties can be solved for a certain class of statistics which we are going to introduce in 2.1 before defining our jackknife and bootstrap in 2.2 and 2.3. This class is sufficiently general to include many statistics of interest.

2.1 Estimators defined by functionals on empirical distributions

For observations $X_1, ..., X_N$ from a stationary process the empirical m-dimensional marginal is

$$\rho_N^m = (N - m + 1)^{-1} \sum_{t=0}^{N-m} \delta_{(X_{t+1}, ..., X_{t+m})} \tag{2.1}$$

where δ_y denotes the point mass at $y \epsilon \mathbb{R}^m$. In this paper we will always consider statistics T_N of the form

$$T_N(X_1, ..., X_N) = T(\rho_N^m) \tag{2.2}$$

with some fixed m and a functional T with values in $\mathbb{R}^q$ defined on the set of all probability measures on $\mathbb{R}^m$ (or a sufficiently rich subset of it). For simplicity of notation we take $q = 1$. Often it is convenient to introduce blocks of observations $Y_t = (X_t, ..., X_{t+m-1})$ and set $n = N - m + 1$. Since $\rho_N^m = n^{-1} \sum_{t=1}^n \delta_{Y_t}$, we are then formally in the case $m = 1$.

In order to illustrate the scope of this class, we give some examples:

i) M, L- and R-estimators of location and scale (see Huber, 1981) are of the form (2.2) with $m = 1$.

ii) Functions of linear statistics:

$T_N = f\left((N-m+1)^{-1}\sum_{t=1}^{N-m+1}\phi(X_t,\dots,X_{t+m-1})\right)$ where

$\phi = (\phi_1,\dots,\phi_k)$ and $f : \mathbb{R}^k \to \mathbb{R}$. This includes the least-square estimator of the parameters of an AR-model and certain versions of the sample correlations.

iii) von Mises' statistics $T_N = N^{-k}\sum_{t_1=1}^{N}..\sum_{t_k=1}^{N}\phi(X_{t_1},\dots,X_{t_k})$ with a symmetric

kernel ϕ.

iv) Statistics defined implicitly as solutions of an equation

$$\sum_{t=1}^{N-m+1}\psi(X_t,...,X_{t+m-1};T_N) = 0.$$

This includes robust estimators for the AR-model (see Künsch, 1984) and maximum likelihood and conditional least squares in Markov processes.

In the context of ARMA-models, Martin and Yohai (1986) considered estimators which are of a slightly more general form than Example iv). They are defined as solutions of

$$\sum_{t=1}^{N}\psi_t(X_t,X_{t-1},\dots,X_1;T_N) = 0$$

where $\psi_t(a_1,a_2,\dots,a_t;\theta) \to \psi(a_1,a_2,\dots;\theta)$ for all $(a_t)\epsilon\mathbb{R}^\infty$. All our procedures and results can be extended to this class.

In order to investigate the asymptotic properties of variance estimators we need an expression for the asymptotic variance of T_N. So we make the following assumptions on the statistics T_N and the underlying stationary process (X_t) :

(A1) T_N converges almost surely to $T(F^m)$ where F^m denotes the marginal distribution of $(X_1, .., X_m)$.

(A2) The influence function $IF(y, F^m) = lim_{\epsilon \downarrow 0} [T((1-\epsilon)F^m + \epsilon\delta_y) - T(F^m)] / \epsilon$ (see Hampel et al., 1986) exists for all $y \epsilon \mathbb{R}^m$.

(A3) $n^{-\frac{1}{2}} \sum_{t=1}^{n} IF(Y_t, F^m)$ is asymptotically normal with mean zero and variance σ^2_{as} where

$$\sigma^2_{as} = \sum_{k=-\infty}^{\infty} E\left[IF(Y_0, F^m) IF(Y_k, F^m)\right].$$

(A4) The remainder term R_N in the linearization $T(\rho_N^m) = T(F^m) + n^{-1} \sum_{t=1}^{n} IF(Y_t, F^m) + R_N$ is of the order $o_p(n^{-\frac{1}{2}})$.

Obviously, (A3) and (A4) imply that $n^{\frac{1}{2}}(T_N - T(F^m))$ is asymptotically normal with mean zero and variance σ^2_{as}. If (X_t) is ergodic, ρ_N^m converges weakly to F^m, so (A1) follows for instance from weak continuity of T. (A3) holds under mixing conditions on (X_t) and moment conditions on $IF(Y_t, F^m)$. For (A4) many techniques from the i.i.d. case can be carried over. In the examples above rigorous conditions can be found in Gastwirth and Rubin (1975) for i), Denker and Keller (1983) for iii), Bustos (1982) and Tjostheim (1986) for iv).

2.2 The jackknife

For the jackknife we delete or downweight blocks of m-tuples in the calculation of the marginal ρ_N^m :

$$\rho_N^{m,(j)} = (n - ||w_n||_1)^{-1} \sum_{t=1}^{n} (1 - w_n(t-j))\, \delta_{Y_t} \tag{2.3}$$

and calculate

$$T_N^{(j)} = T(\rho_N^{m,(j)}) \qquad (j = 0, .., n - \ell). \tag{2.4}$$

The weights $w_n(i)$ are assumed to satisfy $0 \leq w_n(i) \leq 1$ $(i \epsilon \mathbb{Z})$ and $w_n(i) > 0$ iff $1 \leq i \leq \ell$. Here ℓ is the length of the downweighted block, and $\| w_n \|_1 = \sum w_n(i)$. In many cases w_n will be of the form

$$w_n(i) = h\left((i - \frac{1}{2})/\ell\right) \quad (1 \leq i \leq \ell). \tag{2.5}$$

for a function $h : (0,1) \to (0,1)$ which is symmetric about $x = \frac{1}{2}$ and increasing on $(0, \frac{1}{2})$. Choosing for h the indicator function $1_{(0,1)}$ corresponds to simple deletion of blocks. In some cases however there is an advantage by taking h smooth, see Section 3.1.

The jackknife estimate of the variance of T_N is simply the sample variance of the $T_N^{(j)}$'s with a suitable standardization.

$$\begin{aligned}
&\hat{\sigma}^2_{Jack} = (n - \| w_n \|_1)^2 n^{-1} (n - \ell + 1)^{-1} \| w_n \|_2^{-2} \sum_{j=0}^{n-\ell} (T_N^{(j)} - T_N^{(\cdot)})^2 \\
&\text{with} \quad T_N^{(\cdot)} = (n - \ell + 1)^{-1} \sum_{j=0}^{n-\ell} T_N^{(j)} \quad \text{and} \quad \| w_n \|_2^2 = \sum_{i=1}^{\ell} w_n(i)^2.
\end{aligned} \tag{2.6}$$

That this is the right standardization will become clear in 3.l and 4.l. There we will also discuss the choice of ℓ as a function of sample size and strength of dependence of (X_t).

2.3 The bootstrap

In analogy to the i.i.d. case we select blocks of length ℓ at random instead of deleting blocks. Assuming $n = k\ell$ with $k \epsilon \mathbb{N}$ the bootstrap m-dimensional marginal is therefore

$$\rho_N^{m,*} = n^{-1} \sum_{j=1}^{k} \sum_{t=S_j+1}^{S_j+\ell} \delta_{Y_t} \tag{2.7}$$

where $S_1, \ldots, S_k$ are i.i.d. uniform on $\{0, 1, \ldots, n-\ell\}$. We can also write this as

$$\rho_N^{m,*} = n^{-1} \sum_{t=1}^{n} f_t \delta_{Y_t} \quad \text{with} \quad f_t = \#\{j; t-\ell \leq S_j \leq t-1\} \tag{2.8}$$

or as

$$\rho_N^{m,*} = n^{-1} \sum_{t=1}^{n} \delta_{Y_t^*} \tag{2.9}$$

where the k blocks $(Y_1^*, ..., Y_\ell^*), (Y_{\ell+1}^*, ..., Y_{2\ell}^*), ..., (Y_{n-\ell+1}^*, ..., Y_n^*)$ are i.i.d. with distribution $(n-\ell+1)^{-1} \sum_{t=0}^{n-\ell} \delta_{(Y_{t+1}, \ldots, Y_{t+\ell})} = \rho_{Y,n}^{\ell}$. We then form the bootstrap statistics

$$T_N^* = T(\rho_N^{m,*}) \tag{2.10}$$

and approximate the unknown distribution of $T_N - T(F^m)$ by the distribution of $T_N^* - T_N$ where $Y_1, ..., Y_n$ are fixed, but $S_1, ..., S_k$ vary. In particular we use

$$\hat{\sigma}_{Boot}^2 = Var^*(T_N^*) = E^*\left[(T_N^* - E^*[T_N^*])^2\right]. \tag{2.11}$$

Here E^* denotes expectation with respect to $S_1, \ldots, S_k$. Similarly we can estimate quantiles or other quantities of interest. Usually the distribution of $T_N^* - T_N$ and $\hat{\sigma}_{Boot}^2$ have to be evaluated by simulation.

The rationale for our proposal is as follows: The distribution of T_N depends on the unknown distribution F^N of $X_1, ..., X_N$. Even asymptotically, the variance of T_N depends on the distribution of the whole process and not on some finite marginal (see (A3)). Obviously it is impossible to estimate F^N from $X_1, ..., X_N$

without assuming a special structure like independence or a linear model. Using independence Efron's bootstrap estimates F^N by $(\rho_N^1)^{\bigotimes N}$, $\bigotimes$ denoting the product of measures. For $m = 1$ our proposal estimates F^N by $(\rho_N^\ell)^{\bigotimes k}$ and thus coincides with Efron's method if $\ell = 1$. However we will let ℓ tend to infinity as $n \to \infty$ since in this way we asymptotically get all marginals correct.

For $m > 1$ our proposal does not give an estimate of F^N. When we write out the bootstrap sample $(Y_1^*, \ldots, Y_n^*)$ in terms of the original observations (X_t), we obtain $n + k(m-1)$ data points. The reason for this is that we do not want to use observations from different independent blocks in the calculation of $\rho_N^{m,*}$. In this way we reduce the effect of joining independent blocks together.

The bootstrap marginal is by (2.8) an empirical marginal with random weights $n^{-1} f_t$. For t, s not at the border $\mathrm{Cov}(f_t, f_s) = (n - \ell + 1)^{-1} k \max(\ell - \mid t - s \mid, 0) - (n - \ell + 1)^{-2} k \ell^2 \sim \max(1 - \mid t - s \mid / \ell, 0)$. This leads to the following modification suggested by a referee. Take

$$\rho_N^{m,*} = \left(\sum_{t=1}^{n} W(t/\ell) \right)^{-1} \sum_{t=1}^{n} W(t/\ell) \delta_{Y_t} \tag{2.12}$$

where $(W(t))_{t \epsilon \mathbf{R}}$ is a positive stationary process with continuous covariance function $R(t)$, independent of (X_t).

2.4 Other methods

Carlstein (1986) has proposed to use non-overlapping subseries. In our notation this means to calculate a suitable standardization of the sample variance of $(T_\ell^{j\ell})_{j=0,\ldots,k-1}$where

$k = [n/\ell]$ and

$$T_\ell^j = T(\rho_\ell^j), \quad \rho_\ell^j = \ell^{-1} \sum_{t=j+1}^{j+\ell} \delta_{Y_t}. \tag{2.13}$$

Obviously we could also take overlapping subseries, i.e. the sample variance of $(T_\ell^j)_{j=0,\ldots,n-\ell}$. We will compare this method with ours in Remarks 3.1, 3.3, and 4.1,iii) below.

A different approach is to directly estimate the asymptotic variance σ_{as}^2 as given in (A3). We assume that the influence function exists also at ρ_N^m so that we can use $IF(y, \rho_N^m)$ as estimator of $IF(y, F^m)$. We then calculate

$$\hat{\sigma}_{Infl}^2 = n^{-2} \sum_{k=-\ell+1}^{\ell-1} \omega_n(k) \sum_{t=1}^{n-|k|} IF\left(Y_t, \rho_N^m\right) IF\left(Y_{t+|k|}, \rho_N^m\right) \tag{2.14}$$

where the $\omega_n(k)$ are lag weights with $\omega_n(k) \to 1$ for fixed k and $\ell = \ell(n) \to \infty, \ell(n)/n \to 0$. It is well known from spectral analysis (cf. Priestley, 1981, Section 6.2.3.) that such weights are needed to obtain consistency. In Chapters 3 and 4 we will see that the jackknife is asymptotically equivalent to (2.14) for some special weights ω_n.

Another method is to fit a parametric model F_θ to the data and then to use

$$\hat{\sigma}_{par}^2 = \sigma_{as}^2\left(F_{\hat{\theta}}\right)/n. \tag{2.15}$$

However, in complicated models like nonlinear time series the calculation of $\sigma_{as}^2\left(F_{\hat{\theta}}\right)$ may be quite difficult. Often it requires extensive simulations so that jackknifing may be a simpler alternative. Moreover there is always the danger that the parametric model does not hold exactly. Because σ^2_{as} is an infinite sum, $\sigma^2_{as}(F)$ and $\sigma^2_{as}(F_\theta)$ may differ considerably even if all finite-dimensional marginals are close and $IF(Y_t, F^m)$ is bounded. This shows that with dependence parametric methods are even more dangerous than in i.i.d. situations.

3. The arithmetic mean

Here we investigate the properties of our procedures for the arithmetic mean. This corresponds to $m = 1, T(F^1) = \int xF^1(dx) = E[X_t] = \mu$ and $IF(x, F^1) = x - \mu$. This functional is linear and allows explicit calculations of all quantities of interest. At the same time it will be the first step towards a theory for general functionals since by (A4) they can be approximated by an arithmetic mean. The effects of the remainder $\mathrm{R_N}$ in (A4) will be discussed in Chapter 4.

3.1 Relation of the jackknife to spectral estimation

By definitions (2.3) and (2.4), we have for any real number c and $0 \leq j \leq n - \ell$

$$\begin{aligned} T_N^{(j)} &= (n- \| w_n \|_1)^{-1} \sum_{t=1}^{n} (1 - w_n(t-j))\, X_t \\ &= T_N + (n- \| w_n \|_1)^{-1} \left[\| w_n \|_1 \, (T_N - c) - \sum_{t=1}^{n} w_n(t-j)(X_t - c) \right] . \end{aligned} \tag{3.1}$$

In order to obtain a transparent formula for $T_N^{(\cdot)}$, we introduce

$$\alpha_n(t) = (\| w_n \|_1)^{-1} (n - \ell + 1)^{-1} \sum_{j=0}^{n-\ell} w_n(t-j)\, . \tag{3.2}$$

Note that $\alpha_n(t) = (n-\ell+1)^{-1}$ for $\ell \le t \le n-\ell+1$ and $\sum \alpha_n(t) = 1$. Hence

$$\hat{\mu}_n = \sum_{t=1}^{n} \alpha_n(t) X_t \tag{3.3}$$

is an unbiased estimator of μ. It is asymptotically equivalent to the arithmetic mean T_N if $\ell = o(n)$. By putting $c = \hat{\mu}_n$ in (3.1) it follows that

$$\begin{aligned} T_N^{(\cdot)} &= T_N + (n- \| w_n \|_1)^{-1} \| w_n \|_1 \left[(T_N - \hat{\mu}_n) - \sum_{t=1}^{n} \alpha_n(t)(X_t - \hat{\mu}_n)\right] \\ &= T_N + (n- \| w_n \|_1)^{-1} \| w_n \|_1 \, (T_N - \hat{\mu}_n). \end{aligned} \tag{3.4}$$

Hence

$$T_N^{(j)} - T_N^{(\cdot)} = -(n- \| w_n \|_1)^{-1} \sum_{t=1}^{n} w_n(t-j)(X_t - \hat{\mu}_n). \tag{3.5}$$

Remark 3.1. Formulas (3.1) and (3.5) show that the jackknife for the arithmetic mean does not change if we replace $\rho_N^{m,(j)}$ by $\| w_n \|^{-1} \sum w_n(t-j)\delta_{Y_t}$ and adjust the standardization. This means that in this case our procedure differs from Carlstein's (1986) method only by using overlapping blocks and general weights w_n.

In order to make our formula for $\hat{\sigma}^2_{Jack}$ more transparent, we again have to introduce some notation. Let v_n be the convolution of w_n with itself:

$$v_n(k) = \sum_{j=1}^{\ell-|k|} w_n(j) w_n(j+ | k |). \tag{3.6}$$

Note that $v_n(0) = \| w_n \|_2^2$. Furthermore we put

$$\beta_n(t,k) = v_n(k)^{-1}(n-\ell+1)^{-1} \sum_{j=0}^{n-\ell} w_n(t-j) w_n(t+ | k | -j) \quad (| k | < \ell). \tag{3.7}$$

As for $\alpha_n(t)$, we have $\beta_n(t,k) = (n-\ell+1)^{-1}$ for $\ell- | k | \le t \le n-\ell+1$ and $\sum_{t=1}^{n-|k|} \beta_n(t,k) = 1$. Hence

$$\tilde{R}_n(k) = \sum_{t=1}^{n-|k|} \beta_n(t,k)(X_t - \hat{\mu}_n)(X_{t+|k|} - \hat{\mu}_n) \tag{3.8}$$

is an estimate of the covariance $R(k) = E[(X_t - \mu)(X_{t+k} - \mu)]$ which is very similar to the usual sample covariance except that it has a smaller bias.

Now by inserting (3.5) in the definition (2.6) of $\hat{\sigma}^2_{Jack}$ we arrive at the following

<u>Theorem 3.1.</u> In the case of the arithmetic mean we have

$$\hat{\sigma}^2_{Jack} = n^{-1} \sum_{k=-\ell+1}^{\ell-1} v_n(k)/v_n(0)\tilde{R}_n(k). \tag{3.9}$$

In particular, $n\hat{\sigma}^2_{Jack}$ turns out to be a lag weight estimate of the spectral density at zero. If we choose the weights accordingly, it almost coincides with $\hat{\sigma}^2_{Infl}$ from (2.14).

In order to make a precise statement about the bias of $n\hat{\sigma}^2_{Jack}$ as an estimator of σ^2_{as}, we first investigate the bias of $\tilde{R}_n(k)$.

<u>Lemma 3.1.</u> If $\ell = o(n)$ and $\sum |\, j \,||\, R(j) \,| < \infty$, then

i) $\mathrm{Var}(\hat{\mu}_n) = n^{-1} \sum_{j=-\infty}^{\infty} R(j) + O(\ell/n^2)$

ii) $E[\tilde{R}_n(k)] - R(k) = -\mathrm{Var}(\hat{\mu}_n) + O(\ell/n^2) = -n^{-1} \sum R(j) + O(\ell/n^2)$.

The proof is given in the appendix.

Together with (3.9) this gives

$$\begin{aligned} E\left[n\hat{\sigma}^2_{Jack}\right] = \sigma^2_{as} + &\sum_{k=-\ell+1}^{\ell-1} (v_n(k)/v_n(0) - 1)R(k) \\ &- \sum_{|k|\geq \ell} R(k) - n^{-1}\sigma^2_{as} \parallel v_n \parallel_1 /v_n(0) + O(\ell^2/n^2). \end{aligned} \tag{3.10}$$

If $w_n(i)$ is of the form (2.5), then

$$v_n(k) \sim \ell\, h * h(k/\ell) \tag{3.11}$$

The asymptotic bias thus depends on the smoothness of the convolution $h * h$ at zero.

Theorem 3.2. Consider the jackknife of the arithmetic mean with w_n of the form (2.5) and $\ell = \ell(n) \to \infty$. Then

i) If $h(x) = 1_{(0,1)}(x)$, if $\ell = o(n^{1/2})$ and $\sum \mid k \mid\mid R(k) \mid < \infty$, $E\left[n\hat{\sigma}^2_{Jack}\right] - \sigma^2_{as} \sim -\ell^{-1} \sum \mid k \mid R(k)$,

ii) If $h * h$ is twice continuously differentiable around zero, if $\ell = o(n^{\frac{1}{3}})$ and $\sum k^2 \mid R(k) \mid < \infty$, $E\left[n\hat{\sigma}^2_{Jack}\right] - \sigma^2_{as} \sim \ell^{-2} \frac{1}{2}(h*h)''(0)/h*h(0) \sum k^2 R(k)$.

Remark 3.2.

i) Theorem 3.2 demonstrates the advantage of choosing a smooth h.

ii) Theorem 3.2 justifies the choice of the standardization for $\hat{\sigma}^2_{Jack}$ at least for the arithmetic mean. If we multiply $\hat{\sigma}^2_{Jack}$ by $n/(n- \parallel v_n \parallel_1 / v_n(0))$, the fourth term in (3.10) which is due to estimation of μ drops out. Moreover, for $\ell = 1$ our formula would agree with the usual jackknife. However, the effect is small, and the results of the next chapter do not suggest that this factor brings an improvement also for more general statistics.

iii) The condition $\sum |k|^j |R(k)| < \infty$ $(j = 1$ or $2)$ excludes models with long-range dependence. As an example, let us consider the case $R(k) \sim |k|^{-\beta}$, $0 < \beta < 1$. A lengthy, but not difficult calculation gives $E[\hat{\sigma}^2_{Jack}] \sim n^{-1}\ell^{1-\beta} 2(1-\beta)^{-1}(2-\beta)^{-1}$ if $\ell = o(n)$ and $\ell(n) \to \infty$. On the other hand $\mathrm{Var}[T_N] \sim n^{-\beta} 2(1-\beta)^{-1}(2-\beta)^{-1}$. The jackknife thus underestimates the true variance systematically by a factor $(\ell/n)^{1-\beta}$ which tends to zero. In situations with long-range dependence, the only possibility to obtain confidence intervals with asymptotically correct levels seems to be to treat the decay exponent β as an unknown parameter and to estimate it, cf. Hampel et al. (1986), Section 8.1. Beran (1986) has given a procedure which takes the variability of the estimated

β for Gaussian observations into account. Nothing seems to be known about non-Gaussian cases.

iv) Our procedures can be generalized to spatial data $(X_t)_{t\in\mathbb{Z}^2}$ in an obvious way. In that situation the smaller bias of $\widetilde{R}_n(k)$ becomes important. The bias of the usual sample covariance is asymptotically not negligible, see Guyon (1982).

The calculation of $\mathrm{Var}[\widehat{\sigma}^2_{Jack}]$ is lengthy. In the literature there are expressions for the asymptotic variance of a lag weight special estimate, see Priestley (1981), Formula (6.2.113), or Brillinger(1975), Theorems 5.6.2 and 5.9.1. However, there is a small problem with the conditions used by these authors. In the next chapter we want to apply our results not to (X_t) itself, but to $\big(IF(Y_t, F^m)\big)$. It is not clear whether their conditions carry over from the former to the latter. The following theorem uses the strong mixing coefficients $\alpha(k)$ and thus avoids this difficulty.

Theorem 3.3. Consider the jackknife for the arithmetic mean and assume that $E[|X_t|^{6+\delta}] < \infty$ and $\sum k^2\alpha(k)^{\delta/(6+\delta)} < \infty$. If the w_n are of the form (2.5) with $\ell(n) = o(n)$, then

$$\mathrm{Var}(n\widehat{\sigma}^2_{Jack}) \sim \ell n^{-1} 2\sigma^4_{as} \int_{-1}^{1} h * h(x)^2 \mathrm{dx}/h * h(0)^2 \tag{3.12}$$

The proof is given in the appendix.

Corollary 3.1. Under the conditions of Theorem 3.3 the jackknife for the arithmetic mean is consistent if $\ell = o(n)$ and $\ell(n) \to \infty$.

Remark 3.3. If we use disjoint blocks as proposed in Carlstein (1986), then $\mathrm{Var}(n\widehat{\sigma}^2_{Jack}) \sim \ell n^{-1} 2\sigma^4_{as}$. In the case $h(x) = 1_{(0,1)}(x)$ (3.12) becomes $\ell n^{-1} 4\sigma^4_{as}/3$. Since this reduction is substantial, overlapping blocks should be used when one can afford the computations.

From Theorems 3.2 and 3.3 we can determine the optimal order of $\ell(n)$ by equating bias squared and variance. It is in the first case $\ell = O(n^{1/3})$ and in the second case $\ell = O(n^{1/5})$. The mean square error of $n\hat{\sigma}^2_{Jack}$ is then $O(n^{-2/3})$ and $O(n^{-4/5})$ respectively. In order to determine not only the order of $\ell(n)$, but also the constant in front of it, we need to know σ^2_{as} and $\sum |k|^j R(k)$ for $j = 1$ or 2. This is actually more than what we are trying to estimate. Carlstein (1986) proposes a method to overcome this well known dilemma of spectral analysis. He assumes an exponential decay of $\big(R(k)\big)$ and estimates the decay coefficient. Another possibility is to use subjective judgement based on inspection of the sample correlations.

3.2. Validity of the bootstrap approximation

By definition the blocks $(X^*_{j\ell+1}, ..., X^*_{(j+1)\ell})$ (j=0,...,k-1) are i.i.d. with distribution ρ^ℓ_n. This implies that the bootstrapped arithmetic mean T^*_N is of the following form:

$$T^*_N = k^{-1} \sum_{i=1}^{k} U_{n,i} \tag{3.13}$$

where the $U_{n,i}$'s are i.i.d. with

$$P\big[U_{n,i} = (X_{j+1} + ... + X_{j+\ell})/\ell\big] = (n - \ell + 1)^{-1} \quad (j = 0, ..., n - \ell).$$

In particular

$$\begin{aligned} E[T^*_N | X_1, ..., X_N] = E[U_{n,1}] &= (n - \ell + 1)^{-1} \ell^{-1} \sum_{j=0}^{n-\ell} \sum_{t=1}^{\ell} X_{j+t} \\ &= (n - \ell + 1)^{-1} \ell^{-1} \sum_{t=1}^{n} X_t \big(\min(t - 1, n - \ell) - \max(t - \ell, 0) + 1\big) \end{aligned} \tag{3.14}$$

and

$$\mathrm{Var}[T_N^*|X_1,...,X_N] = k^{-1}\mathrm{Var}(U_{n,1}) = k^{-1}(n-\ell+1)^{-1}\ell^{-2}\sum_{j=0}^{n-\ell}\left(\sum_{t=1}^{\ell}(X_{j+t}-E[U_{n,1}])\right)^2 \tag{3.15}$$

Comparing this with (3.2), (3.3) and (3.5) we see that the bootstrap and the jackknife with simple deletion lead to the same variance estimate.

Theorem 3.4. If we choose the weights $w_n(i) = 1 \quad (1 \le i \le \ell)$ for the jackknife, then in the case of the arithmetic mean $\widehat{\sigma}^2_{Boot} = \widehat{\sigma}^2_{Jack}$.

In particular $\widehat{\sigma}^2_{Boot}$ can be calculated without using simulations. For $\mathcal{L}(T_N^* - T_N \mid X_1,...,X_N)$ to be asymptotically a valid approximation of $\mathcal{L}(T_N-\mu)$ the first two moments should be asymptotically equivalent:

(B1) $n\widehat{\sigma}^2_{Boot} \quad \to \quad \sigma^2_{as}$ almost surely

(B2) $n^{1/2}\big(E[T_N^*|X_1,...,X_N]-T_N\big) \quad \to \quad 0$ almost surely.

Theorems 3.2-3.4 show that $E[n\widehat{\sigma}^2_{Boot}] \to \sigma^2_{as}$ and $\mathrm{Var}(n\widehat{\sigma}^2_{Boot}) = O(\ell n^{-1})$. Similarly to Lemma 3.1 i) we can show that $\mathrm{Var}\big(E[T_N^*|X_1,...,X_N]-T_N\big) = O(\ell n^{-2})$. These results imply convergence in probability, but not almost sure convergence. We have to consider fourth moments. If we can show that

$$E\big(n^2/\ell^2\big(n^{1/2}(E[T_N^*|X_1,...,X_N]-T_N)\big)^4\big) \le \text{const.} \tag{3.16}$$

and

$$E\big(n^2/\ell^2(n\widehat{\sigma}^2_{Boot}-E[n\widehat{\sigma}^2_{Boot}])^4\big) \le \text{const.}, \tag{3.17}$$

(B1) and (B2) follow for $\ell = O(n^{1/2-\epsilon})$. The left hand side of (3.16) contains fourth moments. Using the inequality (A.1) of the appendix it can be proved under the conditions of Theorem 3.3. (3.17) involves eighth moments. Similarly to (A.1) we can express an eighth moment by cumulants and estimate these cumulants with the mixing coefficients $\alpha(j)$. However in order to complete the proof, we need an even faster decay of the $\alpha(j)'s$ than in Theorem 3.3.

The distribution of $T_N^* - T_N$ is by (3.13) a k-fold convolution. So except for a very small k direct calculation is not advisable. However (3.13) together with the central limit theorem for triangular arrays suggest that $\mathcal{L}(T_N^*|X_1, ..., X_N)$ will be approximately normal like $\mathcal{L}(T_N)$. For Lindeberg's condition we need an assumption on the fluctuations of partial sums:

(B3) With probability one $\max_{0\leq j\leq n-\ell} \mid \sum_{t=j+1}^{j+\ell} (X_t - \mu)) \mid = o(n^{1/2})$.

The most elegant way to prove (B3) is via strong approximations. If there is a Wiener process W(t) such that with probability one

$$\sum_{s\leq t} (X_s - \mu) - W(t) = o(t^{1/2}), \tag{3.18}$$

(B3) follows for $\ell(n) = O(n^\alpha)$, $\alpha < 1$, from (3.18) by well known properties of the Wiener process. For (3.18) one needs $E|X_i|^p < \infty$ for some $p > 2$ and a quick decay of the strong mixing coefficients, cf. Chapter 7 of Philipp and Stout (1975). If $\ell(n)$ increases slowly, (B3) can be obtained without any restriction on the dependence structure.

<u>Lemma 3.2.</u> (B3) holds if $E|X_t|^p < \infty$ for some $p > 2$ and $\ell(n) = o(n^{1/2-1/p})$.

The proof is given in the appendix.

With the above conditions we can show

<u>Theorem 3.5.</u> If (A1), (A3) and (B1)–(B3) hold, then

$$\sup_x |P[(T_N^* - T_N) \leq x|X_1, ..., X_N] - P[(T_N - \mu) \leq x]| \to 0$$

almost surely.

This is an extension of Theorem 1A of Singh (1981) and of Theorem 2.1 of Bickel and Freedman (1981).

Remark 3.4. For the arithmetic mean we can have approximate equality between $\widehat{\sigma}^2_{Boot}$ and $\widehat{\sigma}^2_{Jack}$ for arbitrary weights of the form (2.5) if we use the proposal (2.12) with $R = h * h$. A rigorous analysis of this method seems however difficult.

A natural question is whether the bootstrap distribution contains more than just the normal approximation. If $\sum \mid E[(X_0 - \mu)(X_t - \mu)(X_s - \mu)] \mid (\mid t \mid + \mid s \mid) < \infty$, a straightforward calculation shows that

$$N^2\{E\left[E[(T^*_N - E[T^*_N \mid X_1, \ldots, X_N])^3 \mid X_1, \ldots, X_N]\right] - E[(T_N - \mu)^3]\} \to 0 \quad (3.19)$$

Hence at least on the average the bootstrap distribution has the correct skewness. This might be surprising because by joining independent blocks we reduce the dependence. But since we are dealing with weak dependence, the main contributions come from short lags which are well approximated by the bootstrap.

4. Nonlinear statistics

4.1 The jackknife

Linearization of T at F^m gives

$$T_N = T(F^m) + n^{-1} \sum_{t=1}^{n} IF(Y_t, F^m) + R_N = L_N + R_N \tag{4.1}$$

$$\begin{aligned} T_N^{(j)} &= T(F^m) + (n - \| w_n \|_1)^{-1} \sum_{t=1}^{n} (1 - w_n(t-j)) IF(Y_t, F^m) + R_N^{(j)} \\ &= L_N^{(j)} + R_N^{(j)}. \end{aligned} \tag{4.2}$$

We want to show that for large n $\hat{\sigma}^2_{Jack}(T_N)$ behaves similarly as $\hat{\sigma}^2_{Jack}(L_N)$. L_N is an arithmetic mean, so the behaviour of $\hat{\sigma}^2_{Jack}(L_N)$ follows from Chapter 3. In particular, under the conditions given there $n\,\hat{\sigma}^2_{Jack}(L_N)$ converges in probability to σ^2_{as}, the asymptotic variance of T_N.

The simplest way is to use results on the asymptotic order of $R_N^{(j)}$ which is in regular cases $O_p(n^{-1})$. Hence we expect $\sum R_N^{(j)2}$ to be $O_p(n^{-1})$. A sufficient condition for this is for instance $\max E[R_N^{(j)2}] = O(n^{-2})$.

Because

$$\sum (R_N^{(j)} - R_N^{(\cdot)})^2 \leq \sum R_N^{(j)2} \tag{4.3}$$

and the standardizing factor in $\hat{\sigma}^2_{Jack}$ is $O(\ell^{-1})$, the following theorem follows from the Cauchy-Schwartz inequality.

<u>Theorem 4.1.</u> If $\sum_{j=0}^{n-\ell} R_N^{(j)2} = O_p(n^{-1})$ and w_n is of the form (2.5), then $n\,\hat{\sigma}^2_{Jack}(T_N) = n\,\hat{\sigma}^2_{Jack}(L_N) + O_p(\ell^{-\frac{1}{2}})$. Hence under the conditions of Theorem 3.3 $n\,\hat{\sigma}^2_{Jack}(T_N)$ converges to σ^2_{as} in probability.

However this result is not completely satisfactory because the contribution of the remainders $R_N^{(j)}$ seems to be of higher order than the difference $n\,\hat{\sigma}^2_{Jack}(L_N) - \sigma^2_{as}$. For this reason we do not check the condition on $R_N^{(j)}$ in concrete examples, but we develop a different approach which gives better results.

In the above argument the inequality (4.3) is too coarse. The remainders $R_N^{(j)}$ typically contain many similar terms which partially cancel if we subtract $R_N^{(\cdot)}$. In order to exploit this without using higher-order expansions, we linearize T at ρ_N^m instead of F^m. Assuming that $IF(y, \rho_N^m)$ exists we write

$$\begin{aligned} T_N^{(j)} &= T_N + (n- \parallel w_n \parallel_1)^{-1} \sum_{t=1}^{n} (1 - w_n(t-j)) IF(Y_t, \rho_N^m) + S_N^{(j)} \\ &= M_N^{(j)} + S_N^{(j)}. \end{aligned} \tag{4.4}$$

Thus we have split $R_N^{(j)}$ into $S_N^{(j)}$ and $M_N^{(j)} - L_N^{(j)}$. First we consider the contribution of $S_N^{(j)}$. It is much smaller than $R_N^{(j)}$ because $\rho_N^{m,(j)}$ and ρ_N^m are much closer than $\rho_N^{m,(j)}$ and F^m. For instance the total variation distance is

$$d_{TV}(\rho_N^{m,(j)}, \rho_N^m) \leq \sum_{t=1}^{n} \mid (n- \parallel w_n \parallel_1)^{-1}(1 - w_n(t-j)) - n^{-1} \mid \leq 2\ell n^{-1}. \tag{4.5}$$

Hence we expect $S_N^{(j)}$ to be at most of the order $O_p(\ell^2 n^{-2})$. With the same arguments as above we can show

<u>Theorem 4.2.</u> If $\sum S_N^{(j)2} = O_p(\ell^4 n^{-3})$, if $\hat{\sigma}^2_{Jack}(M_N) = O_p(n^{-1})$ and w_n is of the form (2.5), then $n\,\hat{\sigma}^2_{Jack}(T_N) = n\,\hat{\sigma}^2_{Jack}(M_N) + O_p(\ell^{3/2} n^{-1})$.

The following lemmas which are proved in the appendix give sufficient conditions for $\sum S_N^{(j)2} = O_p(\ell^4 n^{-3})$. We do not strive for maximal generality, but rather for simplicity of the arguments, see also Remark A.2.

<u>Lemma 4.1.</u> Let T_N be a function of linear statistics like in Example ii) of 2.1. Assume that f is differentiable with $\frac{\partial f}{\partial x_i}$ Lipschitz in a neighbourhood of $\int \phi dF^m$ and that $\phi_1, \cdots, \phi_k$ are bounded. Then $\max_j \mid S_N^{(j)} \mid \leq \text{const.} \ell^2 n^{-2}$.

Lemma 4.2. Let T_N be a von Mises' statistic like in example iii) of 2.l. If the kernel ϕ is bounded, $\max_j \mid S_N^{(j)} \mid \leq \text{const.}\ell^2 n^{-2}$.

Lemma 4.3. Let T_N be like in example iv) of 2.l. Assume that there is a $C \epsilon \mathbb{R}$ and a neighbourhood U_o of $T(F^m)$ such that

i) $\mid \psi(y,\theta) \mid \leq C$ on $\mathbb{R}^m \times U_o$,

ii) $\dot{\psi} = \frac{\partial}{\partial \theta}\psi$ exists and $\mid \dot{\psi}(y,\theta) \mid \leq C$ on $\mathbb{R}^m \times U_o$,

iii) $\mid \dot{\psi}(y,\theta_1) - \dot{\psi}(y,\theta_2) \mid \leq C \mid \theta_1 - \theta_2 \mid$ for $y \epsilon \mathbb{R}^m, \theta_1, \theta_2 \epsilon U_o$,

iv) $\int \dot{\psi}(y, T(F^m)) dF^m(y) \neq 0$.

Then $\max_j \mid S_N^{(j)} \mid \leq \text{const.}\ell^2 n^{-2}$.

We consider next the contribution of $L_N^{(j)} - M_N^{(j)}$. Since $M_N^{(j)}$ is like $L_N^{(j)}$ a weighted average, we are going to use formula (3.9) for $\hat{\sigma}^2_{Jack}(L_N)$ and $\hat{\sigma}^2_{Jack}(M_N)$. Except for the nonstandard version of the sample covariances $\hat{\sigma}^2_{Jack}(M_N)$ equals $\hat{\sigma}^2_{Infl}$. We thus obtain also results on the difference between $\hat{\sigma}^2_{Jack}(L_N)$ and $\hat{\sigma}^2_{Infl.}$. In order to proceed, we need more information on the difference of $IF(y,F^m)$ and $IF(y,\rho_N^m)$.

Lemma 4.4. In the situation and under the conditions of Lemma 4.1 or 4.3 the following condition is satisfied:

(C) There exist bounded functions $h_i : \mathbb{R}^m \to \mathbb{R}$ and random variables $U_{n,i} = O_p(n^{-1/2})$ $(i = 1, \ldots, q)$ such that

$$\sup_y \mid IF(y,F^m) - IF(y,\rho_N^m) - \sum_{i=1}^{q} h_i(y) U_{n,i} \mid = O_p(n^{-1}).$$

The proof is given in the appendix. For von Mises' statistics (C) holds if the kernel ϕ is a linear combination of products of functions of one argument.

Theorem 4.3. If (C) holds and w_n is of the form (2.5), then $n\,\hat{\sigma}^2_{Jack}(M_N) = n\,\hat{\sigma}^2_{Jack}(L_N) + O_p(n^{-1/2}) + O_p(\ell n^{-1})$.

The proof is given in the appendix.

Corollary 4.1: Under the conditions of Theorems 4.2 and 4.3, $n\,\hat{\sigma}^2_{Jack}(T_N) = n\hat{\sigma}^2_{Jack}(L_N) + O_p(\max(\ell^{3/2}n^{-1}, n^{-1/2}))$.

Remarks 4.1.

i) According to Theorem 3.3 and Corollary 4.1, the standard deviation of the jackknife for the linear part dominates all effects of the nonlinearity if $\ell \to \infty$ and $\ell = o(n^{1/2})$. In particular the asymptotically optimal choice of ℓ depends only on the correlations of $IF(Y_t, F^m)$.

ii) Theorems 4.2 and 4.3 show that for $\ell = o(n^{1/3})$ the dominating effect of the nonlinearity is the use of the empirical influence function $IF(y, \rho_N^m)$ instead of $IF(y, F^m)$. Compared to this the difference between $\hat{\sigma}^2_{Jack}$ and $\hat{\sigma}^2_{Infl}$ is negligible. This is of interest also in the classical jackknife with i.i.d. observations and $\ell = 1$ for all n.

iii) An analogue of Theorem 4.1 holds also for Carlstein's (1986) method, but Theorems 4.2 and 4.3 are limited to our procedure. In order to obtain the exact order of the effect of the nonlinearity on Carlstein's method, we consider first i.i.d. observations where in regular situations $E[(T_N - T(F^m))^2] - E[(L_N - T(F^m))^2] = \gamma n^{-2} + o(n^{-2})$ with $\gamma \neq 0$. From this it follows immediately that $E[n\,\widehat{\sigma}^2_{Carl}(T_n)] - E[n\,\widehat{\sigma}^2_{Carl}(L_N)] \sim \gamma\ell^{-1}$. The same behavior is expected to hold under suitable conditions of weak dependence (Theorem 4.1 is not sharp here because of Cauchy-Schwartz). But this means that the nonlinearity of T introduces an additional bias which is of the same order as the bias of the linear part. In particular the optimal choice of ℓ depends also on the nonlinear part. Moreover the effects of the nonlinearity on our procedure are by Corollary 4.1 of smaller order if $\ell = o(n^{2/5})$.

4.2 The bootstrap

Unfortunately the analysis of our bootstrap procedure turns out to be even more difficult than for the standard bootstrap with $\ell = 1$. If we want to generalize Theorem 3.5 to smooth nonlinear functionals T with $m = 1$, we need the distribution of the Kolmogorov-Smirnov distance between $\rho_N^{1,*}$ and ρ_N^1. For $\ell = 1$ it is not too difficult to show that it behaves like the Kolmogorov-Smirnov distance between ρ_N^1 and F^1 for i.i.d. observations,see also Bickel and Freedman (1981), Chapter 4. The difficulty with $\ell > 1$ is that the weights $\rho_N^{1,*}$ attributes to two order statistics become dependent in an uncontrollable way.

The only case which is fairly easy is the one where T is a function of linear statistics as in Example ii) of 2.1. With the delta technique Theorem 3.5 can be generalized to such statistics, cf. Bickel and Freedman (1981), Formula (3.6). Moreover in this situation the bootstrap distribution will usually give a better approximation than the central limit theorem with an estimated variance. The reason for this is that when the bootstrap and the true distribution are in good agreement for the linear statistics $n^{-1}\sum\phi(Y_t)$, then they will be in good agreement for any transformed statistics $f\left(n^{-1}\sum\phi(Y_t)\right)$. However if f is strongly nonlinear, the distribution of $f\left(n^{-1}\sum\phi(Y_t)\right)$ can be very nonnormal.

Theoretical work to confirm this heuristics is now in progress. An empirical confirmation is given by the simulations reported in Table 5 below.

5. Examples

5.1 The trimmed mean

As our first example we estimate the variance of the 40%-trimmed mean in the situation considered by Carlstein (1986). The observations (X_t) come from an AR(1)-process with innovations $\epsilon_t \sim 0.7\mathcal{N}(0,1) + 0.3\mathcal{N}(0,10)$. We calculated $\hat{\sigma}^2_{Jack}$ with weights obtained from the split cosine window

$$w(i) = \begin{cases} \frac{1}{2}(1 - cos(\pi(i - \frac{1}{2})/\ell_1) & 1 \leq i \leq \ell_1 \\ 1 & \ell_1 + 1 \leq i \leq \frac{1}{2}(\ell + 1) \\ w(\ell + 1 - i) & \frac{1}{2}(\ell + 1) \leq i \leq \ell \end{cases} \tag{5.1}$$

and compared it with Carlstein's (1986) method of subseries. Since the differences between the two methods are expected to be larger for small samples, we restricted ourselves to the sample size $n = 100$.

Table 1 shows that for $\beta = 0.2$ the jackknife is a clear improvement over the subseries method whereas for $\beta = 0.8$ the improvement is within the order of magnitude of the random error of the simulations. This can be explained by the different subseries lengths chosen by Carlstein's method. For $n = 100$ the optimal subseries lengths are 3 and 13 for $\beta = 0.2$ and $\beta = 0.8$ respectively. Now for short subseries the nonlinear terms become important and increase the bias of Carlstein's method. For longer subseries the effect of the nonlinear terms disappears. Even then the jackknife should have a smaller variance because it corresponds to using overlapping subseries, but this is hardly visible in our simulations.

5.2 AR-parameters

In the remaining examples we consider the least-squares estimator T_N for the parameters $\theta = (\beta_1, \ldots, \beta_p, \nu)$ of an AR(p)-model

$$X_t = \sum_{j=1}^{p} \beta_j X_{t-j} + \nu + \epsilon_t. \tag{5.2}$$

It is obviously a multivariate function of linear statistics like in Example ii) of 2.l. At the same time it belongs also to the class iv) of 2.1 with $m = p+1$ and multivariate $\psi, \psi_1 = x_p(x_{p+1} - \sum \beta_j x_{p+1-j} - \nu), \ldots, \psi_p = x_1(x_{p+1} - \sum \beta_j x_{p+1-j} - \nu), \psi_{p+1} = x_{p+1} - \sum \beta_j x_{p+1-j} - \nu$. This latter interpretation shows that the influence function for each component of T is a linear combination of $\psi_1, \ldots, \psi_{p+1}$, see Hampel et al. (1986), Formula (4.2.9).

The asymptotic variance of the components of T_N depends on the distribution of the data. If they follow the AR-model (5.2), the psi-function satisfies

$$E_\theta[\psi_j(X_t, \ldots, X_{t+p}; \theta) \mid X_{t+p-1}, X_{t+p-2}, \ldots] \equiv 0. \tag{5.3}$$

This holds then also for the influence function. Therefore $E_\theta[IF_j(Y_t, \theta) IF_k(Y_s, \theta)] = 0$ for all $t \neq s$ and all j, k, cf. Künsch (1984), Section 1.4. By the results of Chapters 3 and 4, the optimal choice for ℓ is $\ell(n) \equiv 1$ provided the model (5.2) holds. If it does not hold, the influence functions are correlated and we have to let ℓ increase with n in order to obtain consistency.

5.2.1 The sunspot numbers

As an example with real data we consider Wolf's sunspot numbers from 1770-1889. Efron and Tibshirani (1986) used the same data set so that we can compare our procedure with theirs. The results are summarized in Table 2 . For $p = 1$ there are two groups of methods. The first one estimates the standard error close to 0.05 and contains methods 1.a, 2.a, 3.a, 4, and 5. The remaining methods estimate the standard error to be 0.035 or 0.036. The first group contains all methods which are based on the assumption that the AR(1)-model holds. However it is well known that the AR(1)-model does not fit at all to the sunspot numbers. We thus trust the second group more. It is also known that an AR(2)-model gives a better fit. With this model the standard error of the AR(1)-coefficient becomes 0.023 which is closer to the value of the second group.

Next we look at the case $p = 2$. The most striking feature is the large difference between the methods 1.a, 2.a, and 3.a on the one hand and methods 4 and 5 on the other hand. Remember that all of these methods should give roughly the same answers if the AR(2)-model were correct. The simulations reported in Table 3 show that these differences cannot be reasonably explained by chance variations. A better explanation is that the AR(2)-model does not hold even though it fits better than AR(1). This is known also from other analyses of the data, see Priestley (1981), Chapter 11. Because of this we again trust the estimates of 1.b, 2.b, and 3.b most.

We also computed the bootstrap histograms for $\hat{\beta}_1$ and $\hat{\beta}_2$. In contrast to the ones obtained by Efron and Tibshirani (1986), they looked very much like normal distributions.

5.2.2 Simulations for the AR(1)-parameters

The results of some simple simulation experiments are shown in Table 4. Note that for $p = 1$ $T_{N,1}$ is just a version of the lag one correlation. First we see that Carlstein's (1986) method is by far the worst in all situations. Because the statistic considered here is highly nonlinear, this method needs much longer subseries in order to achieve a bias comparable to the other methods. This leads then to a much larger standard deviation.

The performance of the other methods depends on the number and the distribution of the observations. In the AR(1)-case the bias of all methods is similar. The parametric method 5 has the smallest variance as was to be expected. In accordance to the theory the variance increases as ℓ increases.

In the MA(1)-case the methods differ also in their bias. The methods 1.a, 2.a, 3,a, and 5 are always biased upwards. Increasing ℓ reduces the size of the bias and changes its sign for $n = 48$. Except in the case of the bootstrap the variance increases with ℓ. From the point of view of mean square error, the estimates with a larger bias but smaller variance are better for the small sample size $n = 48$. There the parametric estimate is by far best although the model is wrong. With the larger sample size $n = 192$ the bias becomes dominant and the nonparametric methods 1.b, 2.b, 3.b with $\ell > 1$ are best. The number of simulations is too small to see differences between the jackknife, the bootstrap and the influence function.

Since the statistic considered is strongly nonlinear, its distribution is not close to the normal. We wanted to see how much the bootstrap distribution picks up of this nonnormality. For the bootstrap distribution one needs more replicates than for the bootstrap variance, see Efron and Tibshirani (1986).We thus restricted ourselves to one situation and took ν to be known.The results of Table 5 show that the bootstrap indeed reflects the nonnormality of the distribution to some extent. How much depends on the block size ℓ. In particular, small $\ell's$ are not good although $\ell = 1$ is optimal for the jackknife.

Appendix: Proofs of results from Chapters 3 and 4

Proof of Lemma 3.1: For i) we write $\hat{\mu}_n - \mu =$ $\sum_{t=1}^{\ell-1} \alpha_n(t)(X_t - \mu) + (n-\ell+1)^{-1} \sum_{t=\ell}^{n-\ell+1} (X_t - \mu) + \sum_{t=n-\ell+2}^{n} \alpha_n(t)(X_t - \mu) =$ $Z_1 + Z_2 + Z_3$ say. By a standard argument $\mathrm{Var}(Z_2) = n^{-1} \sum R(j) + o(\ell n^{-2})$. Furthermore by definition (3.2) $\alpha_n(t) \le (n-\ell+1)^{-1}$.
Hence $\mathrm{Var}(Z_1) = \mathrm{Var}(Z_3) \le (n-\ell+1)^{-2} \sum_{t,s=1}^{\ell-1} \mid R(t-s) \mid = O(\ell n^{-2})$ and $\mid \mathrm{Cov}(Z_1, Z_2) \mid = \mid \mathrm{Cov}(Z_3, Z_2) \mid \le (n-\ell+1)^{-2} \sum_{t=1}^{\ell-1} \sum_{s=\ell}^{n-\ell+1} \mid R(t-s) \mid = O(n^{-2})$.
Finally $\mid \mathrm{Cov}(Z_1, Z_3) \mid \le (n-\ell+2)^{-2} \ell \sum_{|j|>n-2\ell} \mid R(j) \mid = o(\ell n^{-3})$.

For ii) we have by definition $E[\tilde{R}_n(k)] - R(k) = \mathrm{Var}(\hat{\mu}_n) - \sum_{t=1}^{n-|k|} \sum_{s=1}^{n} \beta_n(t,k)$ $\alpha_n(s) \left(R(t-s) + R(t+\mid k \mid -s)\right)$. Now $\beta_n(t,k) = \alpha_n(t)$ for $\ell \le t \le n-\ell+1$ and $\mid \beta_n(t,k) - \alpha_n(t) \mid \le 2(n-\ell+1)^{-1}$. So the assertion follows by similar arguments as above.

Proof of Theorem 3.3: Let $S_n^{(j)} = \sum w_n(t-j)(X_t - \mu)$. In a first step one shows that $\mathrm{Cov}(S_n^{(0)}, S_n^{(j)}) \sim v_n(j)\sigma_{as}^2$ and $\sum \mathrm{Cov}(S_n^{(0)}, S_n^{(j)})^2 \sim \sigma_{as}^4 \sum v_n(j)^2 \sim$ $\ell^3 \sigma_{as}^4 \int h * h(x)^2 dx$. This is done by straightforward calculations using $\sum j^2 \mid R(j) \mid < \infty$ and the form (2.5) of w_n. In a next step one shows that $\sum \mid \mathrm{Cov}(S_n^{(0)2}, S_n^{(j)2}) - 2\mathrm{Cov}(S_n^{(0)}, S_n^{(j)})^2 \mid = o(\ell^3)$. This is done by using the following inequality for $s \leq t \leq u \leq v$ (assuming $\mu = 0$)

$$| E[X_s X_t X_u X_v] - E[X_s X_t]E[X_u X_v] - E[X_s X_u]E[X_t X_v] - E[X_s X_v]E[X_t X_u] | \leq \text{const.}\alpha(\max(t-s, u-t, v-u))^{\delta/(6+\delta)} \quad (A.1)$$

which follows by a repeated application of Theorem 17.2.3 of Ibragimov-Linnik (1971). In a last step one has to show that on the right hand side of (3.5) we may replace $\hat{\mu}_n$ by μ, i.e. $\mathrm{Var}(n\hat{\sigma}^2_{Jack}) \sim n^{-2} \parallel w_n \parallel_2^{-4} \sum \mathrm{Cov}(S_n^{(j)2}, S_n^{(k)2})$. For this we use again (A.1). We leave the details to the reader.

Remark A.1: An alternative proof is obtained by observing that the proof of Theorems 5.6.2 and 5.9.1 by Brillinger (1975) uses only summability of fourth order cumulants. This follows from (A.1) and $\sum k^2 \alpha(k)^{\delta/(6+\delta)} < \infty$. In general summability of $n-th$ order cumulants follows from $\sum k^{n-2}\alpha(k)^{\delta/(2n-2+\delta)} < \infty$ and appropriate moments.

Proof of Lemma 3.2: By the Borel-Cantelli-Lemma $| X_t - \mu | = O(t^{1/p})$ almost surely. Hence $\max | \sum_{t=s+1}^{s+\ell} (X_t - \mu) | = O(n^{1/p}\ell) = o(n^{1/2})$.

Proof of Theorem 3.5: We have to show that the conditional distribution of $n^{1/2}\sigma_{as}^{-1}(T_N^* - T_N)$ given $X_1, \ldots, X_N$ is asymptotically standard normal. By (B2) we may replace T_N by $E[T_N^* \mid X_1, \ldots, X_N]$. By (3.13) T_N^* is a sum of k i.i.d. variables $U_{n,i}$, and by (B1) the conditional variance of $n^{1/2}\sigma_{as}^{-1}T_N^*$ converges to 1. Hence we only have to check the Lindeberg condition which is $E\left[Z_{n,i}^2 1_{[|Z_{n,i}|>\epsilon]}\right] = o(k^{-1})$ where $Z_{n,i} = n^{-1/2}\ell(U_{n,i} - E[U_{n,i}])$. But $P[| Z_{n,i} |> \epsilon] = (n - \ell + 1)^{-1}\#\{j; 0 \leq j \leq n - \ell,$
$| \sum_{t=j+1}^{j+\ell} (X_t - E[U_{n,i}]) |> \epsilon n^{1/2}\}$ which is zero for n large enough by (B2) and (B3).

Proof of Lemma 4.1: A Taylor expansion to the first order shows that

$$S_N^{(j)} = \sum_{i=1}^{k} \left(\frac{\partial f}{\partial x_i} \left(\int \phi \ d\rho_N^m \right) - \frac{\partial f}{\partial x_i}(\tau_{n,j}) \right) \int \phi_i d(\rho_N^{m,(j)} - \rho_N^m)$$

where $\| \tau_{n,j} - \int \phi \ d\rho_N^m \| \leq \| \int \phi \ d(\rho_N^{m,(j)} - \rho_N^m) \|$. By the Hahn decomposition $| \int \phi_i d(\rho_N^{m,(j)} - \rho_N^m) | \leq sup \mid \phi_i(y) \mid d_{TV}(\rho_N^{m,(j)}, \rho_N^m)$. Hence the lemma follows from (4.5) and the assumed Lipschitz continuity of $\frac{\partial f}{\partial x_i}$.

Proof of Lemma 4.2: We have

$$\begin{aligned} S_N^{(j)} &= \sum_{i=2}^{k} \binom{k}{i} \int \phi(x_1, \ldots, x_k) d(\rho_N^{m,(j)} - \rho_N^m)(x_1) \ldots d(\rho_N^{m,(j)} - \rho_N^m)(x_i) \\ &\quad d\rho_N^m(x_{i+1}) \ldots d\rho_N^m(x_k). \end{aligned}$$

The lemma follows therefore directly from the Hahn decomposition of $\rho_N^{m,(j)} - \rho_N^m$.

Proof of Lemma 4.3: By (A1), the ergodic theorem and the assumptions of the lemma $\int \dot{\psi}(y, T_N) d\rho_N^m(y) \neq 0$ for N large enough. Hence by a standard argument (compare Lemma 7.2.l.B of Serfling, 1980) $\int \psi(y,t) d\rho_N^{m,(j)}(y) = 0$ has a solution $T_N^{(j)}$ such that $\max_j \mid T_N^{(j)} - T_N \mid \to 0$. A Taylor expansion to the first order gives

$$\begin{aligned} 0 = \int \psi(y, T_N^{(j)}) d\rho_N^{m,(j)}(y) &= \int \psi(y, T_N) d\rho_N^{m,(j)}(y) \\ &+ \int \dot{\psi}(y, \tau_{j,N}) d\rho_N^{m,(j)}(y)(T_N^{(j)} - T_N). \end{aligned} \tag{A.2}$$

where $\mid \tau_{j,N} - T_N \mid \leq \mid T_N^{(j)} - T_N \mid$. By the assumptions of the lemma

$$\begin{aligned} &\left| \int \dot{\psi}(y, \tau_{j,N}) d\rho_N^{m,(j)}(y) - \int \dot{\psi}(y, T_N) d\rho_N^m(y) \right| \\ \leq &\left| \int \left(\dot{\psi}(y, \tau_{j,N}) - \dot{\psi}(y, T_N) \right) d\rho_N^{m,(j)}(y) \right| + \left| \int \dot{\psi}(y, T_N) d(\rho_N^{m,(j)} - \rho_N^m)(y) \right| \\ \leq &C \left(\mid T_N^{(j)} - T_N \mid + O(\ell n^{-1}) \right) \to 0 \end{aligned} \tag{A.3}$$

uniformly in j. But then we can solve (A.2) for $T_N^{(j)} - T_N$ to obtain

$$\begin{aligned} T_N^{(j)} - T_N &= \left(- \int \dot{\psi}(y, \tau_{j,N}) d\rho_N^{m,(j)}(y) \right)^{-1} \int \psi(y, T_N) d(\rho_N^{m,(j)} - \rho_N^m)(y) \\ &= O(\ell n^{-1}) \text{ uniformly in } j. \end{aligned} \tag{A.4}$$

By the usual formula for the influence function of an M-estimator (see Hampel et al. (1986), Formula (2.3.5)) $S_N^{(j)} = \left((\int \dot{\psi}(y, T_N) d\rho_N^m(y))^{-1} - (\int \dot{\psi}(y, \tau_{j,N}) d\rho_N^{m,(j)}(y))^{-1} \right)$ $\int \psi(y, T_N) d(\rho_N^{m,(j)} - \rho_N^m)(y)$. The lemma follows now by combining (A.3) and (A.4).

Remark A.2: The essential point in the proof of Lemmas 4.1 and 4.3 is the straightforward estimate $\max_j \mid \int g d(\rho_N^{m,(j)} - \rho_N^m) \mid \leq sup \mid g(y) \mid \ell n^{-1}$. However this is in general not optimal and boundedness of g is not really necessary. If w_n corresponds to simple deletion, then $\int g d(\rho_N^{m,(j)} - \rho_N^m) = \ell n^{-1}(n - \ell)^{-1} \sum_{t=1}^{n} (g(Y_t) - E[g(Y_t)]) -$ $(n-\ell)^{-1} \sum_{t=j+1}^{j+\ell} (g(Y_t) - E[g(Y_t)])$. Finding a bound for the maximum of the second term is an analogous problem to checking (B3), and the discussion following (B3) applies also here. But note that for good results we need small rates in the strong approximation (3.l8). For (X_t) i.i.d. this is available (see Csörgö and Revesz, 1981) but not for the dependent case (see Phillip and Stout, 1975). General weights complicate the situation further.

Proof of Lemma 4.4: We only give the proof for M-estimators, the other case being straightforward. By Formula (2.3.5) of Hampel et al. (1986) $IF(y, F^m) = M^{-1} \psi(y, T(F^m))$ and $IF(y, \rho_N^m) = M_N^{-1} \psi(y, T_N)$ where $M = -\int \dot{\psi}(y, T(F^m)) dF^m$ and $M_N = -\int \dot{\psi}(y, T_N) d\rho_N^m$. Arguing like in (A.3) shows that $M_N - M = O_p(n^{-\frac{1}{2}})$ and thus $IF(y, F^m) - IF(y, \rho_N^m) =$
$M_N^{-1} (\psi(y, T_N) - \psi(y, T(F^m))) + \psi(y, T(F^m))(M_N^{-1} - M^{-1}) =$
$M^{-1} \dot{\psi}(y, T(F^m))(T_N - T(F^m)) + \psi(y, T(F^m))(M_N^{-1} - M^{-1}) + O_p(n^{-1})$.

Proof of Theorem 4.3: By Assumption (C) $IF(Y_t, \rho_N^m) IF(Y_s, \rho_N^m) - IF(Y_t, F^m)$ $IF(Y_s, F^m) = \sum_{i=1}^{q} U_{n,i} (h_i(Y_t) IF(Y_s, F^m) + h_i(Y_s) IF(Y_t, F^m)) + O_p(n^{-1})$. Inserting this in the Formula (3.9) for the linear jackknife, we obtain $n \left(\hat{\sigma}^2_{Jack}(M_N) - \hat{\sigma}^2_{Jack}(L_N) \right) =$

$\sum U_{n,i}$ times an estimate of the cross spectral density of $(h_i(Y_t))$ and $(IF(Y_t, F^m))$ at zero $+O_p(\ell n^{-1}) = O_p(n^{-1/2}) + O_p(\ell n^{-1})$.

Acknowledgement:

I wish to thank two referees for their careful reading and helpful comments on earlier versions of this paper which lead to a great improvement in the presentation of our results. I also thank Ed Carlstein for discussions.

References

Beran, J. (1986). *Estimation, testing and prediction for self-similar and related processes.* Ph.D. thesis, ETH Zürich, Nr. 8074.

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

Brillinger, D.R. (1975). *Time Series, Data Analysis and Theory.* Holt, Rinehart and Winston, New York.

Bustos, O. (1982). General M-estimates for contaminated $p-th$ order autoregressive processes: consistency and asymptotic normality. *Z. Wahrsch. verw. Geb.* **59**, 491-504.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14**, 1171-1179.

Csörgö, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics.* Akadémiai Kiadó, Budapest.

Davis, W.W. (1977). Robust interval estimation of the innovation variance of an ARMA-model. *Ann. Statist.* **5**, 700-708.

Denker, M. and Keller, G. (1983). On U-statistics and von Mises' statistics for weakly dependent processes. *Z. Wahrsch. verw. Geb.* **64**, 505-522.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

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.

Freedman, D.A. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models. *Ann. Statist.* **12**, 827-842.

Gastwirth, J.L. and Rubin, H. (1975). The behavior of robust estimators on dependent data. *Ann. Statist.* **3**, 1070-1100.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika* **69**, 95-105.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.

Ibragimov, I.A. and Linnik Yu.V. (1971). *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen.

Künsch, H.R. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.* **12**, 843-863.

Martin, R.D., and Yohai, V.J. (1986). Influence functionals for time series. *Ann. Statist.* **14**, 781-818.

Priestley, M.B. (1981). *Spectral Analysis and Time Series.* 2 vols., Academic Press, London.

Philipp, W., and Stout, W. (1975). Almost sure invariance principles for partial sums of weakly dependent random variables. *Memoirs of the AMS* **161**, .

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

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

Tjøstheim, D. (1986). Estimation in nonlinear time series models. *Stoch. Proc. Appl.* **21**, 251-273.

Tukey, J. (1958). Bias and confidence in not quite large samples. Abstract. *Ann. Math. Statist.* **29**, 614.