Bootstraps for Time Series

Peter Bühlmann
Seminar für Statistik, ETH Zürich, CH-8092 Zürich, Switzerland
buhlmann@stat.math.ethz.ch

July 1999

Abstract

We compare and review block, sieve and local bootstraps for time series and thereby illuminate theoretical facts as well as performance on finite-sample data. Our (re-) view is *selective* with the intention to get a new and fair picture about some particular aspects of bootstrapping time series.

The generality of the block bootstrap is contrasted by sieve bootstraps. We discuss implementational dis-/advantages and argue that two types of sieves outperform the block method, each of them in its own important niche, namely linear and categorical processes, respectively. Local bootstraps, designed for nonparametric smoothing problems, are easy to use and implement but exhibit in some cases low performance.

Key words and phrases. Autoregression, block bootstrap, categorical time series, context algorithm, double bootstrap, linear process, local bootstrap, Markov chain, sieve bootstrap, stationary process.

1 Introduction

Bootstrapping can be viewed as simulating a statistic or statistical procedure from an estimated distribution $\hat{P}_n$ of observed data $X_1, \ldots, X_n$. Under dependence, the construction of $\hat{P}_n$ is more complicated and far less 'natural' than Efron's (1979) breakthrough in the independent set-up. We discuss here mainly block, sieve and local bootstraps, which are all in a certain sense nonparametric and model-free: the purpose is to get a fair picture about strengths and weaknesses of such different time series bootstraps. To do so, we focus on theoretical aspects as well as on real performance for finite sample data. So far, only little attention was paid to an overall perspective when comparing different schemes; and in that respect, our *selective* (re-) view offers also valuable new insights. Our access to the topic and point of view is rather different from Léger, Politis and Romano (1992), Efron and Tibshirani (1993, Chs.8.5–8.6), Shao and Tu (1995, Ch.9), Li and Maddala (1996) or Davison and Hinkley (1997, Ch.8), which are all discussing some aspects of bootstrapping time series. Particularly, we include the recently developed sieve and local bootstraps, the latter being suitable for nonparametric smoothing problems.

Extracting information from data is here formalized with a scalar-, vector- or curve-valued estimator $\hat{\theta}$. Estimation of the sampling distribution of $\hat{\theta}$, or pivotized/studentized versions thereof, is essential for various tasks: drawing statistical inference, comparison of competing estimators and improving them; this includes constructing confidence intervals and tests, measuring efficiency of estimators, estimation of risks for selecting a model or tuning parameters, and bagging [Breiman, 1996] for improving prediction in highly complex models.

With time series data, the task of estimating the distribution of $\hat{\theta}$ is much more difficult than for independent observations and methods based on analytic derivations become very soon extremely unpracticle. Still as a relatively simple example, consider an estimator $\hat{\theta}$ which is asymptotically normally distributed around a finite-dimensional parameter θ of interest: under suitable conditions for the stationary data $X_1, \ldots, X_n$,

$$\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}(0, \sigma_\infty^2) \text{ as } n \to \infty. \tag{1.1}$$

But different from the i.i.d. set-up, the asymptotic variance σ_∞^2 is an infinite-dimensional object which is generally not estimable with convergence rate $1/\sqrt{n}$. For example,

$$\sigma_\infty^2 = \sum_{k=-\infty}^{\infty} \mathrm{Cov}(X_0, X_k), \text{ if } \hat{\theta} = n^{-1} \sum_{t=1}^{n} X_t,$$
$$\sigma_\infty^2 = \sum_{k=-\infty}^{\infty} \mathrm{Cov}(\mathrm{IF}(X_0), \mathrm{IF}(X_k)), \ \mathrm{IF}(x) = \frac{\mathrm{sign}(x-\theta)}{2f(\theta)}, \text{ if } \hat{\theta} = \mathrm{med}(X_1, \ldots, X_n),$$

where for the latter, $\theta = F^{-1}(1/2)$ is the median of the cumulative marginal distribution F of X_t having density f. In case of the mean, the asymptotic variance is the spectral density of the data-generating process at zero [normalized by the factor 2π]; in case of the sample median, the spectral density of the process $(\mathrm{IF}(X_t))_{t\in\mathbb{Z}}$ is involved, i.e., a complicated instantaneous unknown transform of the process $(X_t)_{t\in\mathbb{Z}}$. Particularly in the latter case, it would be very awkward to estimate the unknown density f and hence $\mathrm{IF}(\cdot)$ and finally its spectral density. Bootstraps are able to consistently estimate the distribution of $\sqrt{n}(\hat{\theta}-\theta)$, and also its limiting normal distribution, in an automatic way. Of course, as in the case with independent data, time series bootstraps also offer the advantage of higher order accuracy than estimated normal approximations based on (1.1).

We say that a bootstrap scheme is consistent, or first order accurate, for an $\mathbb{R}^q$-valued estimator $\hat{\theta}$, if

$$\sup_{x\in\mathbb{R}^q} \mathbb{P}^*[a_n(\hat{\theta}^* - \theta^*) \le x] - \mathbb{P}[a_n(\hat{\theta} - \theta) \le x] = o_P(1) \ (n \to \infty), \tag{1.2}$$

for some normalizing sequence $(a_n)_{n\in\mathbb{N}}$. In parametric problems of finite dimension, often $a_n = \sqrt{n}$ and $\sqrt{n}(\hat{\theta}-\theta)$ converges to a normal distribution. The centering value θ^*, which is a constant conditional on the original observations $X_1, \ldots, X_n$, is typically *not* chosen as $\hat{\theta}$ like in Efron's i.i.d. bootstrap: details are given later when specifying particular time series bootstraps. Besides approximating the distribution of $\hat{\theta}$, the aim of interest could also be the bootstrap variance $\mathrm{Var}^*(\hat{\theta}^*)$ aiming for a good variance estimate such that $\mathrm{Var}^*(\hat{\theta}^*)/\mathrm{Var}(\hat{\theta}) \xrightarrow{P} 1$. Or the task of interest may be estimation of bias $b_n = \mathbb{E}[\hat{\theta}] - \theta$ by $b_n^* = \mathbb{E}^*[\hat{\theta}^*] - \theta^*$; the value θ^* is here the same as in (1.2). Of course, bootstrap variance and bias together give then an estimate of the mean squared error $\mathbb{E}[(\hat{\theta} - \theta)^2]$, say for comparing different estimators $\hat{\theta}$ including the choice of tuning parameters, e.g.

bandwidths, and model selection; note that the latter is just a problem about choosing a discrete valued tuning parameter. The consistency in (1.2) or consistent bias, variance and MSE estimation are of course not always true. Bootstrap consistency usually holds when in (1.2), $\hat{\theta}$ is asymptotically normal, often with $a_n = \sqrt{n}$. Then, the accuracy is driven by the accuracy of bootstrap variance estimation

$$a_n^2 \mathrm{Var}^*(\hat{\theta}^*) - a_n^2 \mathrm{Var}(\hat{\theta}).$$

Some of the time series bootstraps do a very good job on this task.

As mentioned already, bootstrap techniques in the independent but also the time series case have the potential for higher order accuracy which is not reflected in (1.2): although not directly for the quantity $a_n(\hat{\theta} - \theta)$ in (1.2), but for studentized versions or when adjusting a confidence interval with BC_a or a double bootstrap calibration.

Accuracy of time series bootstraps can thus be examined on two levels: goodness of first order accuracy in (1.2) which is usually driven by the quality of bootstrap variance, and goodness of second order accuracy for studentized statistics or calibrated confidence intervals. Already the first order approach is challenging when dealing with time series: the limiting variance is generally not estimable with $1/\sqrt{n}$ convergence rate, see e.g. the formulae for σ_∞^2 in the discussion of (1.1), indicating the infinite dimension of the problem. For some finite samples in practice, first order schemes might become equally important as their second order counterpart. A substantial effort is given here to the discussion of first order accuracy, but we include also aspects of second order accuracy.

2 Bootstraps for time series: general remarks

2.1 Model based resampling

A straightforward approach is model based: the dependence structure is modeled explicitly and the resample is drawn from the fitted model. This has been pursued in numerous examples and cases; we only mention Freedman (1984) and Bose (1988) for autoregressive models, Rajarshi (1990) for Markov models and Kreiss and Franke (1992) for ARMA models. Such approaches are, of course, inconsistent if the model used for resampling is misspecified.

2.2 Sieve bootstraps

Sieve bootstraps, based on the idea of sieve approximation [Grenander, 1981], should be viewed as nonparametric schemes: a general process $(X_t)_{t \in \mathbb{Z}} \sim P$ is approximated by a family of (semi-) parametric models

$$\{\mathcal{M}_j; j \in \mathbb{N}\}$$

such that $\cup_{j=1}^{\infty} \mathcal{M}_j$ contains [in some sense] the original process P. The question is of course, which parametric model-family $\{\mathcal{M}_j; j \in \mathbb{N}\}$ one should choose. For the purpose of bootstrapping time series, there are at least two particularly interesting families which we discuss in sections 4 and 6.

The sieve bootstrap in general is as follows. Given is the data $X_1, \ldots, X_n$, and a family of models equipped with a model selection rule

$$X_1, \ldots, X_n \mapsto \hat{\mathcal{M}} \in \cup_{j=1}^{\infty} \mathcal{M}_j,$$

is specified. To be more precise, a model should be understood as a set of probability measures

$$\mathcal{M}_j = \{P_{\eta_j}; \eta_j \text{ in some parameter space } \Theta_j\}. \tag{2.1}$$

Denoting by $\hat{\mathcal{M}} = \mathcal{M}_{\hat{j}}$ for some $\hat{j} \in \mathbb{N}$, we estimate in a second step the unknown parameter $\eta_{\hat{j}}$ in $\hat{\mathcal{M}}$ [as if $\hat{\mathcal{M}}$ would be fixed] so that the estimate for the data generating process is

$$\hat{P}_n = P_{\hat{\eta}_{\hat{j}}}.$$

In a third step, we sample from $\hat{P}_n$,

$$X_1^{*S}, X_2^{*S}, \ldots, X_n^{*S} \sim \hat{P}_n, \tag{2.2}$$

which is the sieve bootstrap sample. Finally, sieve bootstrapping an estimator $\hat{\theta} = h_n(X_1, \ldots, X_n)$, which is a measurable function of the original data $X_1, \ldots, X_n$, is defined with the plug-in rule

$$\hat{\theta}^{*S} = h_n(X_1^{*S}, \ldots, X_n^{*S}). \tag{2.3}$$

Sieve bootstrap schemes are from a practical point of view not too different from the model based approach. The renaming of the bootstrap scheme is primarily due to the fact that the theoretical justification is entirely different, allowing for finite-sample model-misspecification. All what is required is that in the asymptotic limit, as sample size tends to infinity, a correct nonparametric model-specification is obtained. In that sense, sieve bootstraps are robust against model-misspecification.

2.3 Block bootstrap

The block bootstrap tries to mimic the behavior of an estimator $\hat{\theta}$ by i.i.d. resampling of blocks $X_{t+1}, \ldots, X_{t+\ell}$ of consecutive observations: the blocking is used so that within a block, the original time series structure is preserved. Such an idea appears in Hall (1985), but the real birth of the block bootstrap is given by Künsch's (1989) seminal paper, explaining in details how and why such a bootstrap works.

As we will see in section 3, the procedure does *not* yield a reasonable estimate of the distribution of the data-generating stochastic process $(X_t)_{t \in \mathbb{Z}}$. Consequently, the block bootstrapped estimator is not defined by the plug-in rule as in (2.3), an issue which makes the procedure in some cases user-unfriendly. On the other hand, the generation of block-type resamples is very easy and the methodology is general to cope with any suitably regular stationary data-generating process and with many estimators $\hat{\theta}$. The scheme has some similarities to block subsampling, with the important difference that the block bootstrapped estimator $\hat{\theta}^{*B}$ is again evaluated with n bootstrap observations; and not with $m \ll n$ as in subsampling.

2.4 Local bootstraps based on independent resampling

So far, we have not said anything about the statistic $\hat{\theta}$ to be bootstrapped, whose distribution is of interest. The previous procedures, i.e., model-based, sieve- and block bootstrap, all give reasonable answers to a large variety of estimators $\hat{\theta}$ whenever the true data-generating process is an element of the specified model, of the asymptotically specified model or just a general stationary process, respectively. At first sight a bit surprisingly, some bootstraps based on independent resampling can be used for nonparametric estimators $\hat{\theta}$ having slower rate of convergence than $1/\sqrt{n}$, e.g., $\hat{\theta}$ a [kernel] smoother of the conditional expectation $\theta(x) = \mathbb{E}[X_t|X_{t-1} = x]$ of a stationary process. The reason for the consistency of such local bootstraps is the 'whitening by windowing principle', cf. Hart (1995), saying that the distribution of $\hat{\theta}$ remains in first order asymptotics the same as for independent samples.

3 Block bootstrap

3.1 The block bootstrap procedure

Proper application of the block bootstrap scheme involves first an adaptation to the problem. Assume that the statistic $\hat{\theta}$ estimates a functional θ, depending on the m-dimensional marginal distribution of the time series. For example, the lag(1)-correlation $\mathrm{Corr}(X_0, X_1)$ in a stationary time series is a functional of the distribution of (X_t, X_{t+1}), corresponding to $m = 2$. Now, build vectors of consecutive observations

$$Y_t = (X_{t-m+1}, \ldots, X_t),\ t = m, \ldots, n. \tag{3.1}$$

Then construct the block-resampling on the basis of the vectorized observations in (3.1). Build overlapping blocks of consecutive vectors $(Y_m, \ldots, Y_{m+\ell-1}), (Y_{m+1}, \ldots, Y_{m+\ell})$, $\ldots, (Y_{n-\ell+1}, \ldots, Y_n)$, where $\ell \in \mathbb{N}$ is the blocklength parameter. For simplicity, assume first that $n - m + 1 = k\ell$ with $k \in \mathbb{N}$. Then, resample k blocks independently,

$$Y_{S_1+1}, \ldots, Y_{S_1+\ell}, Y_{S_2+1}, \ldots, Y_{S_2+\ell}, \ldots, Y_{S_k+1}, \ldots, Y_{S_k+\ell}, \tag{3.2}$$

where the block-starting points $S_1, \ldots, S_k$ are i.i.d. Uniform$(\{m-1, \ldots, n-\ell\})$. These resampled blocks of m-vectors could be referred to the block bootstrap sample. However, as we will see, the block bootstrapped estimator is not just simply defined by the plug-in rule and the notion of a bootstrap sample is not clear. If $n - m + 1$ is not a multiple of ℓ, we resample $k = [(n - m + 1)/\ell] + 1$ blocks but use only a portion of the k-th block to get $n - m + 1$ resampled vectors in total.

The 'good' definition of the block bootstrapped estimator is not entirely straightforward. The vectorization in (3.1) is typically linked to the estimator in that

$$\hat{\theta} \text{ is symmetric in the vectorized observations } Y_m, \ldots, Y_n.$$

It is often assumed that

$$\hat{\theta} = T(F_n^{(m)}) \tag{3.3}$$

where $F_n^{(m)}(\cdot) = (n-m+1)^{-1}\sum_{t=m}^{n} 1_{[Y_t \le \cdot]}$ is the empirical cumulative distribution function of the m-dimensional marginal distribution of $(X_t)_{t\in\mathbb{Z}}$, and T is a smooth functional.

Example A. For the lag(1)-correlation $\theta = \mathrm{Corr}(X_t, X_{t+1})$, consider the estimator $\hat{\theta} = \hat{R}(1)/\hat{R}(0)$ with $\hat{R}(k) = n^{-1}\sum_{t=1}^{n-k}(X_t - \hat{\mu}_X)(X_{t+k} - \hat{\mu}_X)$ $(k \ge 0)$, $\hat{\mu}_X = n^{-1}\sum_{t=2}^{n} X_t$. This estimator $\hat{\theta}$ is symmetric in $Y_2, \ldots, Y_n$ with $Y_t = (X_{t-1}, X_t)$, i.e., $m = 2$, and it is of the form (3.3).

Example B. The GM-estimators in an AR(p) model can be written in the form (3.3) with $m = p+1$. Besides the Gaussian-MLE, this includes estimators being robust against innovation and lagged-value outliers.

The block bootstrapped estimator is defined as

$$\begin{aligned} \hat{\theta}^{*B} &= T(F_n^{(m)*B}), \\ F_n^{(m)*B}(\cdot) &= (n-m+1)^{-1}\sum_{i=1}^{k}\sum_{t=S_i+1}^{S_i+\ell} 1_{[Y_t \le \cdot]}. \end{aligned} \tag{3.4}$$

Generally $\mathbb{E}^{*B}[\hat{\theta}^{*B}] \ne \hat{\theta}$, and the choice for θ^* in (1.2) is often $\mathbb{E}^{*B}[\hat{\theta}^{*B}]$ [say if $\hat{\theta}$ is asymptotically unbiased for θ].

This definition of the block bootstrapped estimator, given by Künsch (1989), can be interpreted as

$$\begin{aligned} \hat{\theta}^{*B} &= g_{n-m+1}(Y_{S_1+1}, \ldots, Y_{S_1+\ell}, Y_{S_2+1}, \ldots, Y_{S_2+\ell}, \ldots, Y_{S_k+1}, \ldots, Y_{S_k+\ell}), \\ \hat{\theta} &= g_{n-m+1}(Y_m, \ldots, Y_n), \end{aligned}$$

saying that it employs a plug-in rule based on the vectorized observations. In particular, the block bootstrapped estimator is defined with values occurring only in the set of the original vectorized observations. This would *not* be the case without the vectorization step in (3.1). Figure 3.1 illustrates the artifact of the naive bootstrap using $m = 1$ instead of the correct $m = 2$, say in Example A. The most striking defects with the naive block bootstrap sample are the newly created scatter plot points within the rectangles in the upper left and the lower right corner. A naively block bootstrapped estimator which uses the plug-in rule in conjunction with the naive block bootstrap sample [e.g., for the autocorrelation in Example A] is then quite strongly affected by these newly created and bad points. As mentioned already, this artifact is not present with the block bootstrap definition in (3.4) based on the vectorized observations.

For the block bootstrap procedure, at least two difficulties remain to be answered in a case by case manner.

(1) Redesigning the *computation* of $\hat{\theta}^{*B}$, which can become practically very inconvenient.

(2) Vectorization as in (3.1) is not always appropriate. For example, the MA-parameter in an MA(1) model or the spectral density of a stationary process depend on the entire distribution of the process, corresponding to $m = \infty$.

Whenever problems (1) and/or (2) become too awkward, an ad-hoc solution is to neglect the vectorization step in (3.1) and work with the naive-block bootstrap [using $m = 1$]. Then, a substantial efficiency loss of the method has generally to paid. Proposals for solving problem (2), mainly in case of spectral density estimation, have been given by Politis and Romano (1992) and Bühlmann and Künsch (1995).

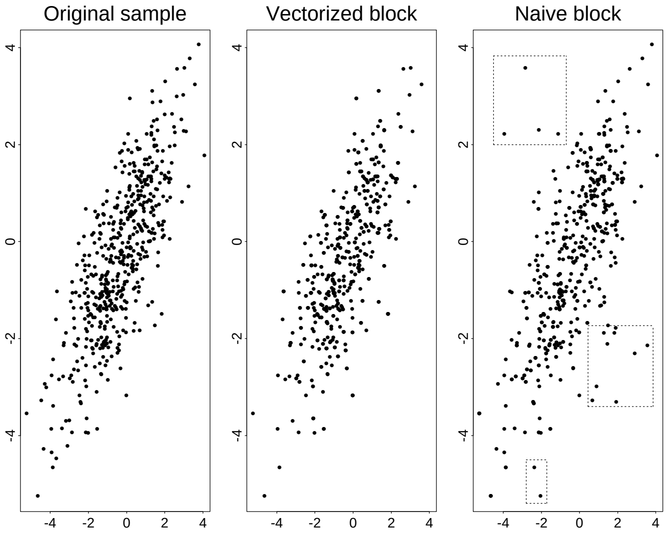

Figure 3.1: Lag(1) scatter plots of re-/samples of size $n = 512$. Left panel: original sample (X_{t-1}, X_t), $t = 2, \ldots, n$. Middle panel: block bootstrap sample Y_{S_i+j}, $i = 1, \ldots, k = 64, j = 1, \ldots, \ell = 8$ from (3.2) with $m = 2$. Right panel: naive block bootstrap sample $(X_{t-1}^{*nB}, X_t^{*nB})$, $t = 2, \ldots, n$, where X_t^{*nB} is the sequentially t-th value in (3.2) with $m = 1, k = 64, \ell = 8$; the points within the rectangles [and others] are not in the plot of the left panel.

3.2 Accuracy

We consider first estimation of the asymptotic variance of $\hat{\theta}$ by the block bootstrap. Künsch (1989) showed that

$$\mathbb{E}[(n\mathrm{Var}^{*B}(\hat{\theta}^{*B}) - n\mathrm{Var}(\hat{\theta}))^2] \sim const.n^{-2/3}, \tag{3.5}$$

achieved with the rate-optimal blocklength $\ell = const.n^{1/3}$. The essential assumptions for this are that $n\mathrm{Var}(\hat{\theta})$ converges to a non-degenerate limiting variance, that T in (3.3) is sufficiently smooth and some mixing conditions on the stationary data-generating process $(X_t)_{t\in\mathbb{Z}}$. A bit surprisingly, the rate $n^{-2/3}$ does *not* depend on the 'degree of dependence', say how fast autocorrelations, or more general mixing coefficients, decay as separation lags increase. In particular, even when autocovariances and mixing coefficients decay exponentially fast, the MSE-rate is still $n^{-2/3}$. Thus, the block bootstrap variance estimate is not rate-adaptive with respect to dependence properties of the underlying process. An explanation of this non-adaptivity was already given by Künsch (1989): the block bootstrap variance estimate is asymptotically equivalent to a lag-window spectral density estimator at the origin with triangular window,

$$n\mathrm{Var}^{*B}(\hat{\theta}^{*B}) \approx \sum_{k=-\ell}^{\ell} (1 - \frac{|k|}{\ell})\hat{R}_{\mathrm{IF}}(k), \tag{3.6}$$

where $\hat{R}_{\mathrm{IF}}(k)$ is the empirical covariance of $(\mathrm{IF}(Y_t; F^{(m)}))_{t=m}^{n}$ at lag k with $\mathrm{IF}(\cdot; F^{(m)})$ the influence function of the estimator at the true underlying m-dimensional marginal distribution $F^{(m)}$. But the triangular form of the window makes it impossible to improve upon the $n^{-2/3}$ MSE-rate, except when the underlying process would be i.i.d. and blocklength $\ell = 1$ [or any fixed, non-increasing number] would be chosen.

For constructing confidence regions, Götze and Künsch (1996) showed that the distribution of a suitably defined studentized version of $\hat{\theta}$ can be approximated by the block bootstrap with accuracy close to $O_P(n^{-2/3})$, using a blocklength $\ell = const.n^{1/3}$. [1] As in variance estimation, the rate of accuracy cannot be improved for time series having geometrically fast dependence properties. For finite samples, the method often behaves erratically and causes problems due to inaccuracy of finite sample variance estimates. Götze and Künsch (1996) also justify a modification of Efron's (1987) BC_a correction. Double block bootstrapping for say a correction of a first order bootstrap confidence region is not easily [if not even im-] possible.

3.3 Choosing a blocklength ℓ

An optimal blocklength depends on at least three things: the data-generating process, the statistic to be bootstrapped and the purpose for which the bootstrap is used, e.g., bias, variance or distribution estimation.

Consider first block bootstrap variance estimation for an estimator $\hat{\theta}$ of the form (3.3). Then,

$$\hat{\theta} \approx (n-m+1)^{-1} \sum_{t=m}^{n} \mathrm{IF}(Y_t; F^{(m)}), \tag{3.7}$$

where $\mathrm{IF}(\cdot; F^{(m)})$ is the influence function as in (3.8). Based on this linearization, formula (3.6) follows and can be rewritten as

$$n\mathrm{Var}^{*B}(\hat{\theta}^{*B}) \approx 2\pi \hat{f}_{\mathrm{IF}}(0), \tag{3.8}$$

where $\hat{f}_{\mathrm{IF}}(\lambda)$ $(0 \le \lambda \le \pi)$ is a triangular window spectral density estimator at frequency λ with bandwidth ℓ^{-1}, based on the influence functions $(\mathrm{IF}(Y_t; F^{(m)}))_{t=m}^{n}$. The blocklength has thus the interesting interpretation as an inverse bandwidth in spectral density estimation. It implies that the asymptotically MSE optimal blocklength for variance estimation is

$$\ell_{opt} = const.n^{1/3}.$$

Bühlmann and Künsch (1999) propose estimation of ℓ_{opt} [i.e., for the constant in the expression above] by an iterative plug-in scheme for optimal local bandwidth choice in spectral density estimation, using the asymptotic equivalence in (3.8).

Regarding block bootstrap bias estimation, consider the case where

$$\theta = H(\mu_f),\ \mu_f = \mathbb{E}[f(X_1, \ldots, X_m)],\ f: \mathbb{R}^{dm} \to \mathbb{R}^q,\ X_t \in \mathbb{R}^d$$
$$\hat{\theta} = H(\hat{\mu}_f),\ \hat{\mu}_f = (n-m+1)^{-1} \sum_{t=m}^{n} f(X_{t-m+1}, \ldots, X_t)$$

[1]This rate can be improved to come close to $O_P(n^{-3/4})$ by using a variance estimate for studentizing which takes negative values with positive probability.

with $H : \mathbb{R}^q \to \mathbb{R}$ a smooth function. Then, the bias $\mathbb{E}[\hat{\theta}] - \theta$ can be estimated by

$$\mathbb{E}^{*B}[H(\hat{\mu}_f^{*B})] - H(\mathbb{E}^{*B}[\hat{\mu}_f^{*B}]).$$

Due to block edge effects, this is better than subtracting $H(\hat{\mu})$. Lahiri (1999) shows that the asymptotic MSE optimal blocklength for bias and variance estimation is the same; estimated blocklengths for variance estimation can thus be used for bias estimation as well.

A method which is more general, and also applicable for choosing an optimal blocklength ℓ for distribution estimation, was proposed by Hall, Horowitz and Jing (1995). They consider the performance of the block bootstrap with different blocklengths for subsamples of size $m \ll n$ yielding an optimal blocklength for subsample size m. The optimal estimated blocklength is then derived with a Richardson extrapolation adjusting to the original sample size n. The method needs a specification of the subsample size m which appears to be less critical than selecting a blocklength. Such subsampling techniques are very general but may not be very efficient; in particular, when the estimator $\hat{\theta}$ is highly nonlinear so that performance on a subsample can be very poor.

Lahiri (1996b) proposes a block-jackknife method for estimating the variance and an extrapolation for estimating the bias of the block bootstrap estimate. The method is again very general.

Automatic choice of the blocklength is at least as difficult as selection of a bandwidth-type tuning parameter in the context of time series. Even worse, consider formula (3.8) which exhibits an equivalence to a bandwidth selection problem: but this is only asymptotically true, since the linearization in (3.7) can have a substantial effect for finite sample size. Furthermore, the blocklength ℓ has no practically relevant interpretation and diagnostic tools are so far undeveloped.

3.4 Range of applicability

The block bootstrap is designed to work for general stationary data generating processes $(X_t)_{t \in \mathbb{Z}}$, typically with $X_t \in \mathbb{R}^d$ $(d \geq 1)$ or also taking values in a categorical space. When restricting to short range dependent processes, the block bootstrap has been theoretically justified in many circumstances: for example for estimators as in (3.3) with smooth T. Some references are given in section 9. In case where the observations have a heavy tailed marginal distribution, Lahiri (1995) shows that block bootstrapping with resampling size $m \ll n$ works for $\hat{\theta} = \overline{X}_n$.

Under long-range dependence, some theory and modifications are worked out in case where $\hat{\theta} = \overline{X}_n$. Lahiri (1993) shows that the block bootstrap is consistent whenever $\overline{X}_n$ has a normal limiting distribution and the bootstrapped statistic is corrected with a factor depending on the typically unknown rate of convergence, e.g. on the self-similarity parameter in self-similar processes. In case where $\overline{X}_n$ has a non-normal limit due to long-range dependence, Hall, Jing and Lahiri (1998) show consistency of a modified block-subsampling procedure.

4 AR-sieve bootstrap for stationary linear time series.

We refer to a linear, invertible time series if it allows an autoregressive representation of order infinity [AR(∞)],

$$X_t - \mu_X = \sum_{j=1}^{\infty} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t \ (t \in \mathbb{Z}), \tag{4.1}$$

where $\mu_X = \mathbb{E}[X_t]$, $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an innovation sequence of i.i.d. random variables with $\mathbb{E}[\varepsilon_t] = 0$ and ε_t independent of $\{X_s; s < t\}$; additional regularity conditions for moments of ε_t and summability of $(\phi_j)_{j\in\mathbb{N}}$ have to be made for a proper definition of an AR(∞).

4.1 The bootstrap procedure

The AR-sieve approximation is as follows. An autoregressive order $\hat{p}$ is chosen, for example with the AIC criterion. The remaining parameter of interest $\eta_{\hat{p}}$ in the AR($\hat{p}$) model is semiparametric [see also formula (2.1)],

$$\eta_{\hat{p}} = (\mu_X, (\phi_1, \ldots, \phi_{\hat{p}}), F_\varepsilon),$$

with F_ε the distribution of the i.i.d. innovations ε_t. The parameter estimate $\hat{\eta}_{\hat{p}}$ is chosen as follows,

$$\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t,$$

$(\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}})$ by the Yule Walker method,

$$\hat{F}_\varepsilon(x) = \mathbb{P}[\varepsilon_t \le x] = (n-\hat{p})^{-1} \sum_{t=\hat{p}+1}^{n} 1_{[R_t - \overline{R}_. \le x]},\ R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j},$$

with $\overline{R}_.$ the mean of the available residuals R_t.

The estimated autoregressive process of order $\hat{p}$ having distribution $\hat{P}_{n;AR} = P_{\hat{\eta}_{\hat{p}}}$ in the notation of section 2.2, is given by

$$X_t^{*AR-S} - \hat{\mu}_X = \sum_{j=1}^{\hat{p}} \hat{\phi}_j (X_{t-j}^{*AR-S} - \hat{\mu}_X) + \varepsilon_t^* \ (t \in \mathbb{Z}), \tag{4.2}$$

with $(\varepsilon_t^*)_{t\in\mathbb{Z}}$ an i.i.d. innovation sequence having marginal distribution $\varepsilon_t^* \sim \hat{F}_\varepsilon$. The AR-sieve bootstrap sample is then a finite sample of size n from the process in (4.2). The AR-sieve bootstrapped estimator $\hat{\theta}^{*AR-S}$ is constructed as in (2.3). This kind of bootstrap was introduced by Kreiss (1992) and further analyzed by Bühlmann (1997), Bickel and Bühlmann (1999) and Choi and Hall (1999).

In (1.2), the parameter of interest θ is a functional of the true underlying process $(X_t)_{t\in\mathbb{Z}} \sim P$ and θ^{*AR-S} is the same functional evaluated at the estimated $\hat{P}_{n;AR} = \hat{P}_{\hat{\eta}_{\hat{p}}}$ which generates the bootstrapped process in (4.2).

Example A [continued]. For the lag(1)-correlation estimator, $\theta^{*AR-S} = \mathrm{Corr}^{*AR-S}(X_t^{*AR-S}, X_{t+1}^{*AR-S})$.

Note that in general $\mathbb{E}^{*AR-S}[\hat{\theta}^{*AR-S}] \neq \theta^{*AR-S}$. The computation of θ^{*AR-S} can be done with a fast Monte Carlo evaluation as follows.

(1) Generate one very long realization $X_1^{*AR-S}, \ldots, X_m^{*AR-S}$ with m much bigger than n.

(2) Use $\hat{\theta}_m^{*AR-S} = h_m(X_1^{*AR-S}, \ldots, X_m^{*AR-S})$ as a Monte Carlo approximation of θ^{*AR-S}.

The justification of the approximation in step (2) is given by formula (1.2) saying that $\hat{\theta}_m^{*AR-S}$ converges to θ^{*AR-S} with rate $a_m^{-1} \ll a_n^{-1}$ [assuming $a_n(\hat{\theta}_n - \theta)$ converges to a non-degenerate distribution].

4.2 Accuracy

Within the class of linear invertible time series as defined in (4.1), the AR-sieve bootstrap is known to have high accuracy: theoretical and practical studies show that it usually outperforms the more general block bootstrap from section 3. In Bühlmann (1997) it is shown that for $\hat{\theta} = \overline{X}_n = n^{-1}\sum_{t=1}^{n} X_t$ and when using an approximating autoregressive order p given by the AIC criterion,

$$n\mathrm{Var}^{*AR-S}(\overline{X}_n^{*AR-S}) - n\mathrm{Var}(\overline{X}_n) = O_P(n^{-(v-2)/(2v)})$$

if the true autoregressive parameters $(\phi_j)_{j\in\mathbb{N}}$ decay like $\phi_j \le const.j^{-v}$ $(v > 2)$. In particular, if the ϕ_j's decay exponentially fast, then

$$n\mathrm{Var}^{*AR-S}(\overline{X}_n^{*AR-S}) - n\mathrm{Var}(\overline{X}_n) = O_P(n^{-1/2+\kappa}) \text{ for any } \kappa > 0. \tag{4.3}$$

The two results show that the method *adapts automatically* to the decay of the underlying dependence structure, a very desirable feature which is not present with the block bootstrap, see (3.5). These adaptivity results are not only asymptotically relevant but can be well seen in finite sample simulations, see section 5.2.

The AR-sieve bootstrap is not only very accurate for variance estimation: Choi and Hall (1999) show some second order property for constructing confidence regions. They propose to calibrate an obtained first order region by double bootstrapping, based on ideas dating back to Hall (1986), Beran (1987) and Loh (1987). Consider construction of a two-sided confidence interval which covers θ with probability $1-\alpha$. A first order interval is given by $[\hat{\theta} - \hat{r}_{1-\alpha/2}, \hat{\theta} - \hat{r}_{\alpha/2}]$, where

$$\hat{r}_\alpha \text{ is the } \alpha\text{-quantile of } \hat{\theta}^{*AR-S} - \theta^{*AR-S} \text{ conditional on } X_1, \ldots, X_n.$$

Now consider an additive correction of the original nominal coverage level by using the double bootstrap. Based on $X_1^{*AR-S}, \ldots, X_n^{*AR-S}$, run the AR-sieve bootstrap to obtain $X_1^{**AR-S}, \ldots, X_n^{**AR-S}$. Now,

$$\hat{r}_\alpha^{*AR-S} \text{ is the } \alpha\text{-quantile of } \hat{\theta}^{**AR-S} - \theta^{**AR-S}, \text{ conditional on } X_1^{*AR-S}, \ldots, X_n^{*AR-S}.$$

Put

$$\hat{a}(1-q) = \mathbb{P}^{*AR-S}[\hat{\theta}^{*AR-S} - \hat{r}_{1-q/2}^{*AR-S} \le \theta^{*AR-S} \le \hat{\theta}^{*AR-S} - \hat{r}_{q/2}^{*AR-S}], \tag{4.4}$$

measuring actual coverage on nominal level $1-q$ [for the second level bootstrap based on the first level bootstrapped data; θ^{*AR-S} is a constant depending only on $X_1, \ldots, X_n$]. Then define

$$\hat{s}_{1-\alpha} = \hat{a}^{-1}(1-\alpha), \text{ the } (1-\alpha)\text{-quantile of } \hat{a} \text{ viewed as a cdf,}$$

which corrects the nominal coverage level $1-\alpha$ to $\hat{s}_{1-\alpha}$. See also Figure 6.4. Now use

$$[\hat{\theta} - \hat{r}_{\{1-(1-\hat{s}_{1-\alpha})/2\}}, \hat{\theta} - \hat{r}_{\{(1-\hat{s}_{1-\alpha})/2\}}] \tag{4.5}$$

as a two-sided, double bootstrap confidence interval for θ with nominal coverage level $1-\alpha$. As shown by Choi and Hall(1999), this interval is second order correct. Importantly, studentization which can be very inefficient in practice, isn't necessary. Instead, second order accuracy is achieved with the double bootstrap, a methodology which generally makes sense with any reasonable sieve-bootstrap scheme. Choi and Hall (1999) report from a simulation study that this second order interval can bring very substantial improvement in some cases and is 'never' significantly worse than the first order construction.

4.3 Choosing the approximating autoregressive order

We propose the AR-sieve approximation in conjunction with the minimum AIC model selection procedure. Shibata (1980) has shown optimality of the AIC for prediction in AR(∞) models. Moreover, (4.3) and its preceding formula which are both based on AIC explain why the criterion is a good choice for variance estimation of $\hat{\theta} = \overline{X}_n$. Since the true model from (4.1) is of infinite autoregressive order, consistency of a model selection scheme does not make much sense.

The optimal autoregressive order generally depends, similar to the optimal blocklength ℓ for the block bootstrap, on the true underlying process, the statistic to be bootstrapped and the purpose for what the bootstrap is used. The AIC criterion automatically selects higher orders for more strongly dependent models; nothing is known how to adapt the order in the AR-sieve approximation to the statistic to be bootstrapped or to the different cases of bootstrap variance- or distribution-estimation.

The tuning element of the AR-sieve bootstrap, namely the selection of an AR-model, has a nice interpretation and allows for diagnostic checks, including graphical procedures. This is in contrast to bandwidth-type tuning parameters, as the blocklength in section 3.3, which have no good interpretation and are not easy to 'back-test' on the data. Last, our experience is that the choice of an approximating autoregressive order, and thus of a model, is quite *insensitive* with respect to the performance of the AR-sieve bootstrap, as long as the chosen order is reasonably good.

4.4 Range of applicability

The AR-sieve bootstrap heavily relies on the crucial assumption that the data $X_1, \ldots, X_n$ comes from an AR(∞)-process as in (4.1). Consistency as in (1.2) for $\hat{\theta}$ being a smooth function of means is given in Bühlmann (1997). The AR(∞) representation includes as the most interesting class the ARMA-models

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{k=1}^{q} \psi_j \varepsilon_{t-k} + \varepsilon_t \ (t \in \mathbb{Z}),$$

with invertible generating MA-polynomial, i.e., $\Psi(z) = 1 + \sum_{k=1}^{q} \psi_k z^k$ ($z \in \mathbb{C}$) has its root outside the unit disk $\{z \in \mathbb{C}; |z| \leq 1\}$; here $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d. innovation sequence and a few additional regularity conditions, being standard in ARMA model theory, have to be made.

Interestingly, as discussed in Bickel and Bühlmann (1996, 1997), the closure [with respect to certain metrics] of the class of linear stationary processes, and also of the class of AR(∞) processes as in (4.1), is surprisingly large. Roughly speaking, to any possibly nonlinear stationary process, there is another process in the closure of linear processes having *exactly the same sample path* with probability greater than $1/e \approx 0.36$. In some sense, this is good news for AR-sieve bootstrapping, saying that the method should also work reasonably good, even if the data is not coming from an AR(∞) source. On the other hand, the fact that the closure is large makes it very delicate or impossible, even when using an infinite sample of data, to test the hypothesis about linearity or AR(∞) representation of the generating process.

5 Block and AR-sieve bootstrap in action

5.1 Total ozone series from Arosa

We study here the world's longest series of total ozone monthly measurements from Arosa, Switzerland, during the period 1926–1997. It is an important source to assess the ozone depletion in the northern mid-latitude hemisphere. The measurements are currently performed by the Swiss Meteorological Institute. The homogenized data set is available from http://www.lapeth.ethz.ch/doc/totozon.html. The raw monthly measurements $\{O_t\}_t$ exhibit big seasonal effects, which can be explained very well. Assuming fixed monthly effects β_i ($i = 1, \ldots, 12$) with $\sum_{i=1}^{12} \beta_i = 0$, we remove them by preliminary smoothing with a running mean $X_t = \sum_{i=-6}^{6} c_i O_{t-i}$ with $c_i = 1/12$ ($i = -5, \ldots, 5$) and $c_i = 1/24$ ($i = -6, 6$). Figure 5.1 displays the filtered data $\{X_t\}_{t=1}^{n}$ with $n = 814$ on the Dobson scale. One main interest is the study of a possibly varying mean trend: an estimate thereof is shown in Figure 5.1. An additional question is about the ozone variability around a varying trend whose estimate is also given in Figure 5.1. We use here time series bootstraps to assess statistical accuracy of these trend and variability smoothers and to answer the questions whether trend and/or variability change significantly over time.

We consider the 'basis' model

$$X_t = m(t/n) + s(t/n)Z_t,\ \ t = 1, \ldots, n = 814,$$

where $m(\cdot)$ and $s(\cdot)$ are smooth mean and scale functions from $[0,1] \to \mathbb{R}$ and $\mathbb{R}^+$, respectively. Moreover, $(Z_t)_{t\in\mathbb{Z}}$ is a stationary process with $\mathbb{E}[Z_t] = 0$ and $\mathrm{Var}(Z_t) = 1$. What is shown in Figure 5.1 are estimates of $m(\cdot)$ and $s(\cdot)$, defined as follows. For the mean function,

$$\hat{m}(x) = \sum_{t=1}^{n} K(\frac{x - t/n}{h})X_t,\ \ 0 < x < 1,$$

where K is the standard Gaussian kernel and bandwidth is $h = 0.024$. For the scale function, build the transformed values

$$\log((X_t - \hat{m}(t/n))^2) \approx \mathbb{E}[\log(Z_t^2)] + \log(s^2(t/n)) + V_t,\ \ t = 1, \ldots, n,$$

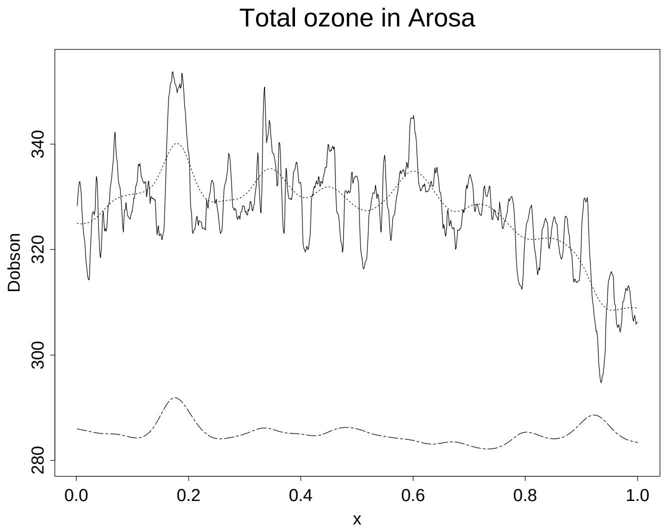

Figure 5.1: Total ozone measurements given by the solid line, mean trend smoother given by the dotted line; magnitude of smoother for changing variability given by the dashed line at the bottom. Time from January 1927 to June 1997 is rescaled to $(0, 1]$.

where $V_t = \log(Z_t^2) - \mathbb{E}[\log(Z_t^2)]$. Now use the same kernel estimator as above applied to $\log((X_t - \hat{m}(t/n))^2)$, estimating $\mathbb{E}[\log(Z_t^2)] + \log(s^2(t/n))$. Transforming back by exponentiating and estimating $\exp(\mathbb{E}[\log(Z_t^2)])$ by matching estimated second moments [using that $\mathrm{Var}(Z_t) = 1$] yields the curve estimate $\hat{s}(\cdot)$.

In the sequel we test the two hypotheses H_1: $m(\cdot)$ is constant, and H_2: $s(\cdot)$ is constant. To do so, we apply some bootstraps to the residual process $\hat{Z}_t = (X_t - \hat{m}(t/n))/\hat{s}(t/n)$ yielding Z_t^* $(t = 1, \ldots, n)$. Bootstrapping from the null-distribution is then done as

$$X_t^{*H_1} = \hat{\mu} + \hat{s}(t/n) Z_t^* \ (t = 1, \ldots, n) \text{ for } H_1,$$

where $\hat{\mu} = n^{-1} \sum_{t=1}^{n} X_t$, and

$$X_t^{*H_2} = \hat{m}(t/n) + \hat{\sigma} Z_t^* \ (t = 1, \ldots, n) \text{ for } H_2,$$

where $\hat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} (X_t - \hat{m}(t/n))^2$. Using the plug-in principle for bootstrapping $\hat{m}(\cdot)$ and $\hat{s}(\cdot)$, inference under the hypotheses can then be done with

$$\hat{m}^{*H_1}(\cdot) \text{ based on } X_1^{*H_1}, \ldots, X_n^{*H_1} \text{ for } H_1, \tag{5.1}$$

$$\hat{s}^{*H_2}(\cdot) \text{ based on } X_1^{*H_2}, \ldots, X_n^{*H_2} \text{ for } H_2. \tag{5.2}$$

The construction of the resampled noise process Z_t^* $(t = 1, \ldots, n)$, being the same for either hypotheses, is done with the AR-sieve and block bootstrap: the former with AIC estimated order 29, the latter with the blocklengths $\ell = 9 \approx n^{1/3}$ and $\hat{\ell} = 25$ which is the estimate from Bühlmann and Künsch (1999) when the statistic of interest would be the arithmetic mean, see section 3.3. Figures 5.2 and 5.3 show the estimates $\hat{m}(\cdot)$ and $\hat{s}(\cdot)$ together with 19 bootstrap replicates each from the estimates in (5.1) and (5.2), respectively.

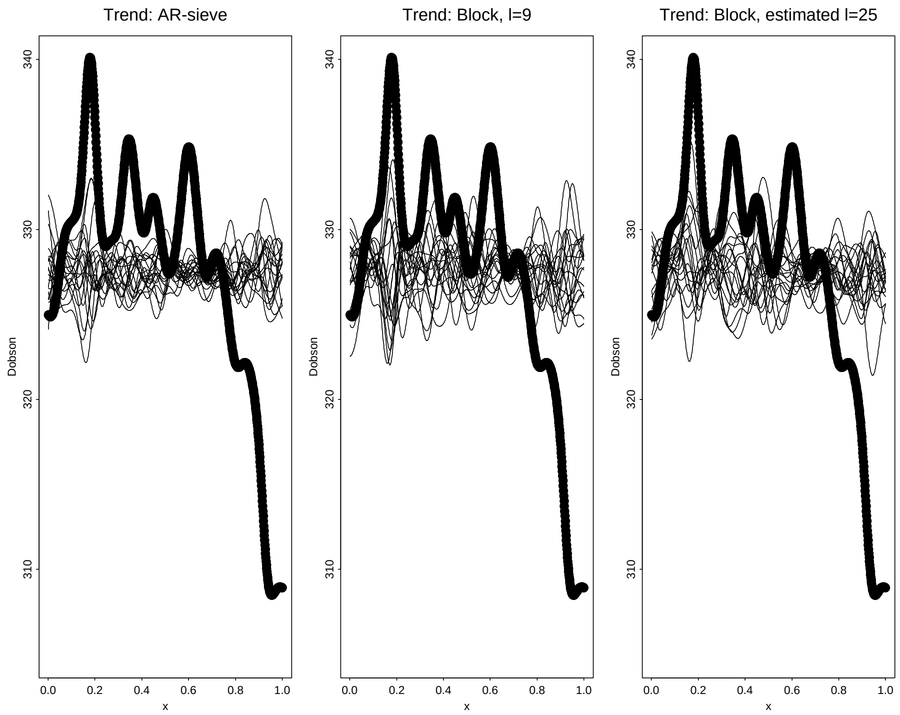

Figure 5.2: Mean trend estimates. 19 bootstrapped estimators $\hat{m}^{*H_1}(\cdot)$ under the hypothesis H_1 with constant trend are displayed by the fine lines, the estimator $\hat{m}(\cdot)$ based on original data is indicated with the bold points. AR-sieve bootstrap with AIC estimated order 29, block bootstrap with $\ell = 9$ and estimated $\hat{\ell} = 25$.

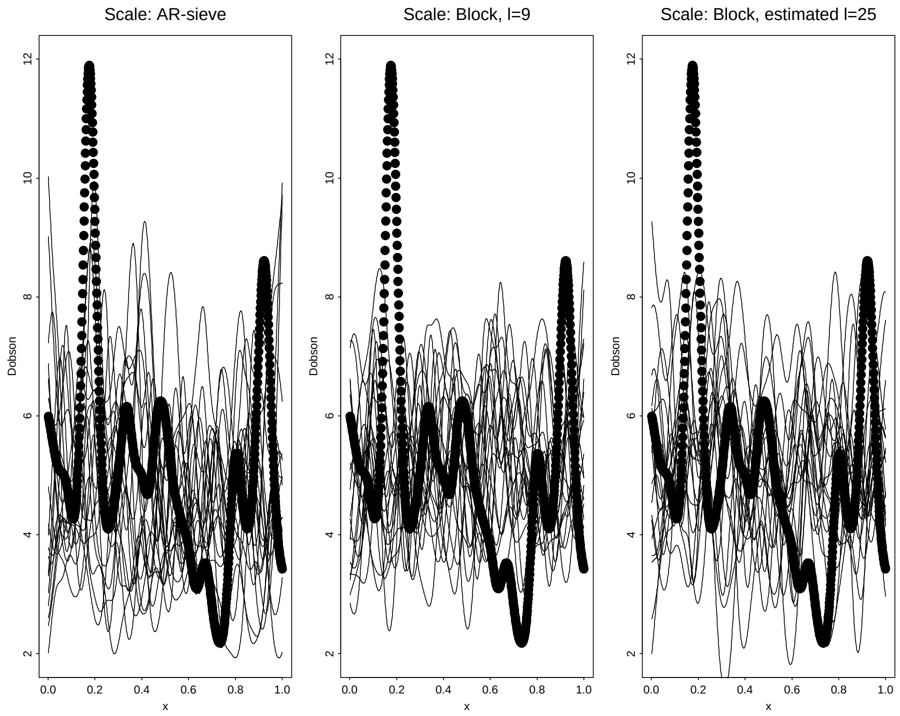

Figure 5.3: Estimates for scale. 19 bootstrapped estimators $\hat{s}^{*H_2}(\cdot)$ under the hypothesis H_2 with constant scale are displayed by the fine lines, the estimator $\hat{s}(\cdot)$ based on original data is indicated with the bold points. AR-sieve bootstrap with AIC estimated order 29, block bootstrap with $\ell = 9$ and estimated $\hat{\ell} = 25$.

They display the '1 out of 19 graphical rule' from Brillinger (1997), asking whether the original estimates $\hat{m}(\cdot)$ and $\hat{s}(\cdot)$ are the most extreme among a set of twenty curves, corresponding to a 5% significance level for testing. Of course, a more formal construction of pointwise or even simultaneous acceptance regions for say two-sided testing of H_1 and H_2 could be given.

All three bootstrap methods lead to similar conclusions which increases confidence about the appropriateness of the graphics in Figures 5.2 and 5.3: it is very valuable to have both, the AR-sieve and block bootstrap as a tool in a practical example. Regarding the mean trend, there is clear evidence that it is changing, actually decreasing, with progressing time. Looking at the scale or variability around the mean trend, there is weak evidence of changing scale, particularly at $x = 0.176$, corresponding in real time to October 1940, and secondary also at $x = 0.923$, corresponding to March 1992.

5.2 AR-sieve versus block bootstrap for simulated series

For comparison of the two bootstraps we consider simulation experiments with two different processes but the statistic being in both cases the sample median $\hat{\theta} = \mathrm{med}(X_1, \ldots, X_n)$. The sample sizes are $n = 512$. Furthermore, the tuning parameters are chosen by the minimal AIC criterion for the AR-sieve; and $\ell = 8 = n^{1/3}$, according to the optimal asymptotic rate for variance estimation, and $\hat{\ell}$ from Bühlmann and Künsch (1999) for the block bootstrap as indicated in section 3.3.

For the first experiment, consider the linear ARMA(1,1) process

$$X_t = -0.8X_{t-1} - 0.5\varepsilon_{t-1} + \varepsilon_t, \tag{5.3}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d sequence, $\varepsilon_t \sim t_6$ independent from $\{X_s; s < t\}$. This model is representable as an AR(∞) process as in (4.1).

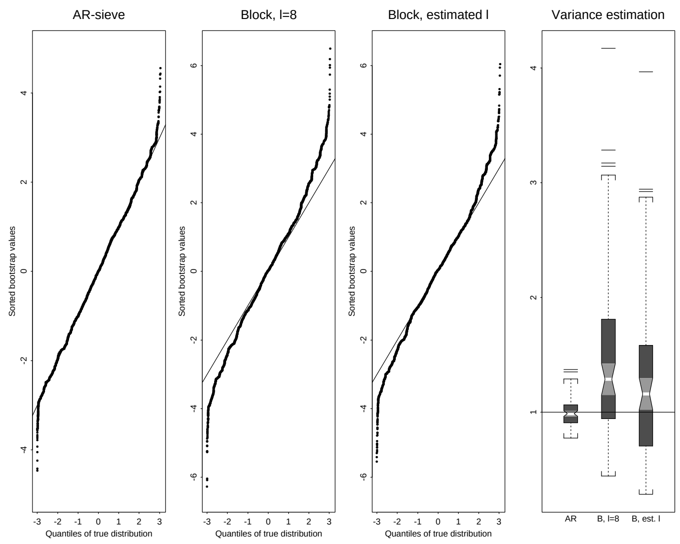

Figure 5.4: Linear model (5.3), $n = 512$: bootstrap distribution and variance estimation of $(\hat{\theta} - \mathbb{E}[\hat{\theta}])/\sigma_n$ by $(\hat{\theta}^* - \mathbb{E}^*[\hat{\theta}^*])/\sigma_n$ for $\hat{\theta} = \mathrm{med}(X_1, \ldots, X_n)$ $[\sigma_n = (\mathrm{Var}(\hat{\theta}))^{1/2}]$. Three left panels: QQ-plots with target indicated by the line. Right panel: boxplots with target indicated by the horizontal line. 100 simulation runs, 500 bootstrap replicates per simulation run.

Figure 5.4 displays the quality of bootstrap approximations in model (5.3). The AR-sieve bootstrap outperforms the block bootstrap very clearly. Estimation of the blocklength improves a bit on the fixed blocklength $\ell = 8 = n^{1/3}$. The fact about the better performance of the AR-sieve bootstrap is not so surprising: Bühlmann (1997) and Hall and Choi(1999) report a general better performance of the AR-sieve method whenever the true underlying process is representable as an AR(∞) as in (4.1). The result here indicates quantitatively the gain in case of a true underlying process which is not a finite order AR-model and hence not an element of the approximating sieve [for any finite sample size]. As noted already in Bühlmann (1997), the gain of the AR-sieve bootstrap is usually more substantial if the autocovariances of the process exhibit some damped pseudo-periodic decay, which is true for the model in (5.3): note that this is a feature which can be graphically diagnosed by looking at estimated autocovariances.

The second experiment is with a nonlinear exponential AR(2) process with heteroscedastic innovations,

$$X_t = (0.5 + 0.9\exp(-X_{t-1}^2))X_{t-1} - (0.8 - 1.8\exp(-X_{t-1}^2))X_{t-2} + \sigma_t\varepsilon_t,$$

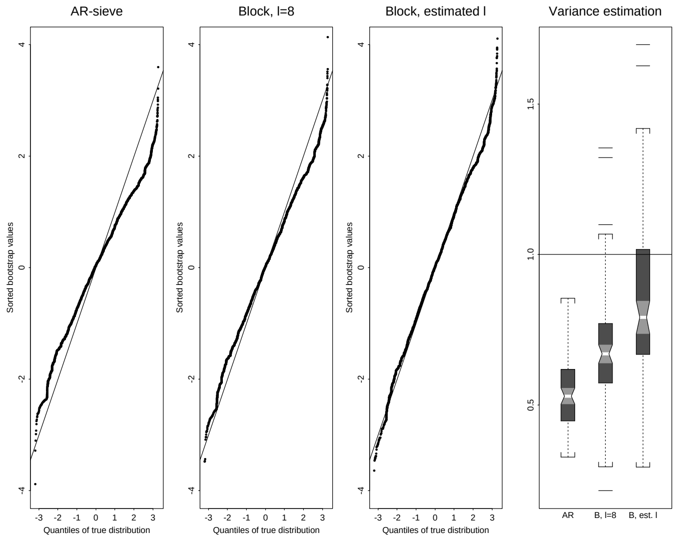

Figure 5.5: Nonlinear model (5.4), $n = 512$: bootstrap distribution and variance estimation of $(\hat{\theta} - \mathbb{E}[\hat{\theta}])/\sigma_n$ by $(\hat{\theta}^* - \mathbb{E}^*[\hat{\theta}^*])/\sigma_n$ for $\hat{\theta} = \mathrm{med}(X_1, \ldots, X_n)$ $[\sigma_n = (\mathrm{Var}(\hat{\theta}))^{1/2}]$. Three left panels: QQ-plots with target indicated by the line. Right panel: boxplots with target indicated by the horizontal line. 100 simulation runs, 500 bootstrap replicates per simulation run.

$$\sigma_t^2 = 0.5 + 0.1X_{t-1}^2 + 0.05\sigma_{t-1}^2 1_{[X_{t-1} \le 0]} + 0.5\exp(-\sigma_{t-1}^2)1_{[X_{t-1}>0]}, \tag{5.4}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d sequence, $\varepsilon_t \sim t_6/\sqrt{1.5}$ independent from $\{X_s; s < t\}$. This process is not representable as an AR(∞) as in (4.1).

Figure 5.5 displays the quality of bootstrap approximations in model (5.4). The AR-sieve bootstrap, which is not asymptotically consistent due to the nonlinearity of the model in (5.4) exhibits a clear bias; and the block bootstrap is superior. As in the linear case (5.3) from above, using the estimated blocklength $\hat{\ell}$ improves upon the fixed blocklength $\ell = 8 = n^{1/3}$. The value of the results in Figure 5.5 is again to get a quantitative idea of the gain when using the block bootstrap.

For other ['weakly'] nonlinear models, the AR-sieve scheme was found to be competitive or even better than the block bootstrap. But the difficulty is to check from data the strength of nonlinearity or the closeness to an AR(∞) representation of the underlying process.

6 VLMC-sieve bootstrap for stationary categorical time series.

Sieve approximation is also successful for general stationary categorical processes $(X_t)_{t\in\mathbb{Z}}$ with values X_t in a categorical, finite space $\mathcal{X}$. The difficulty is the construction of a sieve being flexible enough to cover general processes, but allowing a parsimonious parameterization so that reasonable statistical accuracy can be maintained. We propose the sieve of so-called variable length Markov chains [VLMC] for approximating the $\mathcal{X}$-valued time series $(X_t)_{t\in\mathbb{Z}}$.

An $\mathcal{X}$-valued, stationary VLMC $(X_t)_{t\in\mathbb{Z}}$ is characterized as a Markov chain of high order whose time-homogeneous transition probabilities are depending on a *variable* number ℓ of lagged values,

$$\mathbb{P}[X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots] = \mathbb{P}[X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-\ell} = x_{t-\ell}]$$

for all x_t, where $\ell = \ell(x_{t-1}, x_{t-2}, \ldots)$ is itself a function of the past. If $\ell(x_{t-1}, x_{t-2}, \ldots) \equiv p$ for all $x_{t-1}, x_{t-2}, \ldots$, we obtain the full Markov chain model of order p. For variable $\ell(\cdot)$ with $\sup\{\ell(x_{t-1}, x_{t-2}, \ldots);\ x_{t-1}, x_{t-2}, \ldots\} = p$, we have an embedding [full] Markov chain of order p, but with an additional *structure of a variable length memory*, implying that some transition probabilities of the embedding Markov chain are lumped together. A VLMC can be displayed as a graphical tree model.

Example C. $\mathcal{X} = \{0, 1\}$, variable length memory bounded by $p = 3$.
The function

$$\ell(x_{-\infty}^0) = \begin{cases} 1, & \text{if } x_0 = 0, x_{-\infty}^{-1} \text{ arbitrary} \\ 3, & \text{if } x_0 = 1, x_{-1} = 0, x_{-\infty}^{-2} \text{ arbitrary} \\ 2 & \text{if } x_0 = 1, x_{-1} = 1, x_{-\infty}^{-2} \text{ arbitrary} \end{cases}$$

can be represented by the tree τ_ℓ on the left hand side in Figure 6.1. A 'growing to the left' sub-branch represents the symbol 0 and vice versa for the symbol 1. The state space is given by the terminal nodes $\{0, 100, 101, 11\}$ of the tree [read top down].

Example D. $\mathcal{X} = \{0, 1, 2, 3\}$, variable length memory bounded by $p = 2$.
The function

$$\ell(x_{-\infty}^0) = \begin{cases} 1, & \text{if } x_0 \in \{0, 1, 2\}, x_{-\infty}^{-1} \text{ arbitrary} \\ 1, & \text{if } x_0 = 3,\ x_{-1} \in \{0, 1, 2\}, x_{-\infty}^{-2} \text{ arbitrary} \\ 2 & \text{if } x_0 = 3, x_{-1} = 3, x_{-\infty}^{-2} \text{ arbitrary} \end{cases}$$

can be represented by the tree τ_ℓ on the right hand side in Figure 6.1. The state space

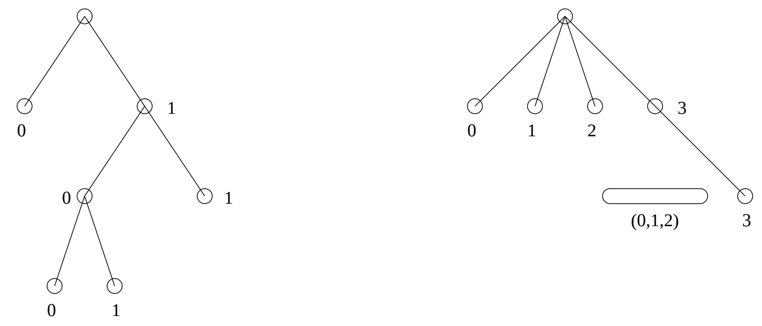

Figure 6.1: Tree representations of the variable length memories in Examples C and D.

$\{0, 1, 2, 3, 33\}$ is again given by the tree: note that the state 3 is an internal node. An alternative representation of the state 3 is given by the round-edged rectangle symbolizing the absent nodes 0,1 and 2 in depth 2, which can be thought as a completion of the tree with nodes lumped together to the state 3.

Fitting a VLMC involves a version of the tree structured context algorithm [Rissanen, 1983] for estimating the variable length memory, described by the function $\ell : \mathcal{X}^\infty \to \cup_{m=0}^{\infty}\mathcal{X}^m$, and the set of transition probabilities $\eta_{\ell(\cdot)}$ in the VLMC with memory given by $\ell(\cdot)$. The exact description of the algorithm as used here can be found in Bühlmann and Wyner (1999). In the notation of section 2.2, the semiparametric model is $\mathcal{M}_{\ell(\cdot)}$ with elements $P_{\eta_{\ell(\cdot)}}$ [which are VLMC's]. Estimation of $\ell(\cdot)$ is a highly complex model selection problem; due to the extremely large number of possible models, a natural tree hierarchy is employed. The context algorithm yields a consistent estimate $\hat{P}_{n;VLMC} = P_{\hat{\eta}_{\hat{\ell}(\cdot)}}$ for the distribution of suitably regular processes which are not necessarily a VLMC, see Bühlmann and Wyner (1999) and Ferrari (1999).

The construction of the VLMC-sieve bootstrap is as follows. Resample

$$X_1^{*VLMC-S}, \ldots, X_n^{*VLMC-S} \sim \hat{P}_{n;VLMC}. \tag{6.1}$$

We then proceed as in (2.3).

6.1 Accuracy and range of applicability

For variance estimation, rate adaptivity with respect to the decay of dependence holds for the VLMC-sieve bootstrap: if the mixing coefficients decay exponentially fast as separation lags increase, and if the data generating process is suitably regular,

$$n\mathrm{Var}^{*VLMC-S}(\hat{\theta}^{*VLMC-S}) - n\mathrm{Var}(\hat{\theta}) = O_P(\log(n)^6/\sqrt{n}), \tag{6.2}$$

where $\hat{\theta} = (n-m+1)^{-1}\sum_{t=m}^{n} f(X_{t-m+1}, \ldots, X_t)$ with $f : \mathcal{X}^m \to \mathbb{R}$ ($m \in \mathbb{N}$). The result describes an adaptation in a *fully data-driven* way to the case of exponential decaying covariances. Note that the process $(X_t)_{t\in\mathbb{Z}}$ is generally not a VLMC.

Double VLMC bootstrapping is potentially possible and construction of a calibrated confidence interval can be done analogously to (4.5), aiming for higher order coverage properties.

The VLMC-sieve bootstrap is designed to be consistent for data-generating stationary categorical processes which are short range dependent, e.g. with 'reasonable' decay of mixing coefficients. Consistency as in (1.2) then holds for general estimators of the form (3.3) defined in section 3.1. More details are given in Bühlmann and Wyner (1999).

6.2 Tuning parameter selection

The tuning parameter of the VLMC-sieve bootstrap is a so-called cutoff value which has to be specified in the tree structured context algorithm for fitting a VLMC. It characterizes a specific tree pruning procedure which is used for computationally efficient selection of a VLMC model-structure. Data-driven selection of this cutoff aiming for minimizing Kullback-Leibler distance or other prediction risks is proposed in Bühlmann (1999) with resampling from a 'hyper'-VLMC model. The method can be adapted to estimate an optimal cutoff for bootstrapping.

6.3 VLMC-sieve versus block bootstrap for simulated series

The differences in variance estimation between (3.5) and (6.2) can be well seen in finite-sample problems, and the gain of the VLMC-sieve bootstrap is often substantial.

We consider a simulated example, where

$$X_t = 1_{[Y_t>0]},\ Y_t = 0.8Y_{t-1} + \varepsilon_t,\ (t \in \mathbb{Z}),$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d. innovation sequence, $\varepsilon_t \sim t_6$ independent from $\{Y_s; s < t\}$. Of interest is the stationary binary process $(X_t)_{t\in\mathbb{Z}}$ whose memory, describing the structure of X_t given $X_{t-1}, X_{t-2}, \ldots$ is non-sparse and infinitely long: thus, a priori, we do not give any advantage to the method of VLMC-sieve approximation.

We consider two different estimators

(S1) $\hat{\theta} = n^{-1}\sum_{t=1}^{n} X_t$,

(S2) $\hat{\theta} =$ VLMC-estimator for the probability of the five-tuple (1,1,1,1,1).

The estimator (S1) is linear and structurally very simple, whereas (S2) is a complicated function of the data, involving a tree-structured model. VLMC-sieve bootstrapping for (S1) and (S2) is according to the plug-in rule in (2.3). Block bootstrapping for (S1) requires no vectorization step. The estimator (S2) is an example where vectorization with $p \geq 5$ would be appropriate, but redesigning the computation with $p \geq 5$ is merely impossible: the VLMC-estimator, whose input is a categorical time series, involves a complicated tree model selection explaining also why $p \geq 5$ is typically unknown; and the probability estimate for the 5-tuple of 1's is a complicated function of parsimoniously estimated transition probabilities. The only feasible way, which we follow, for block bootstrapping (S2) is to neglect the vectorization step and work with $p = 1$. The sample sizes are $n = 128$ for (S1) and $n = 512$ for (S2).

The VLMC-sieve bootstrap is run with different cutoff tuning parameters, the block bootstrap with the blocklength $\ell = 5 \approx n^{1/3}$ for $n = 128$ and $\ell = 8 = n^{1/3}$ for $n = 512$, and for (S1) also with the estimated $\hat{\ell}$ from Bühlmann and Künsch (1999) as indicated in section 3.3. Figure 6.2 displays the quality of distribution estimation for the estimator (S1). For a whole range of cutoff tuning parameters, the VLMC-sieve bootstrap is better than the block bootstrap, and for the latter using the estimated blocklength $\hat{\ell}$ improves a bit upon the fixed blocklength $\ell = 5 \approx n^{1/3}$. Figure 6.3 displays the quality of distribution and variance estimation for (S2). Due to computational expenses when bootstrapping the complicated estimator (S2) we only ran the procedures with one 'standard' tuning parameter, each: cutoff $\chi^2_{1;0.95}/2 = 1.92$ and $\ell = 8$ [estimation of ℓ for the complicated estimator (S2) is very difficult]. Also here, the VLMC-sieve is better than the block bootstrap; the VLMC method shows a few large outliers for variance estimation which indicates a small chance that the VLMC-sieve approximation can be bad for the complicated and thus potentially 'unstable' estimator (S2).

Figures 6.2 and 6.3 are representative for other situations with exponentially decaying dependence structure: generally, the VLMC sieve-bootstrap is superior to the block-bootstrap, the latter being also more sensible to the specification of the blocklength parameter. In practice, an insensitive procedure to the choice of tuning parameters is highly desirable.

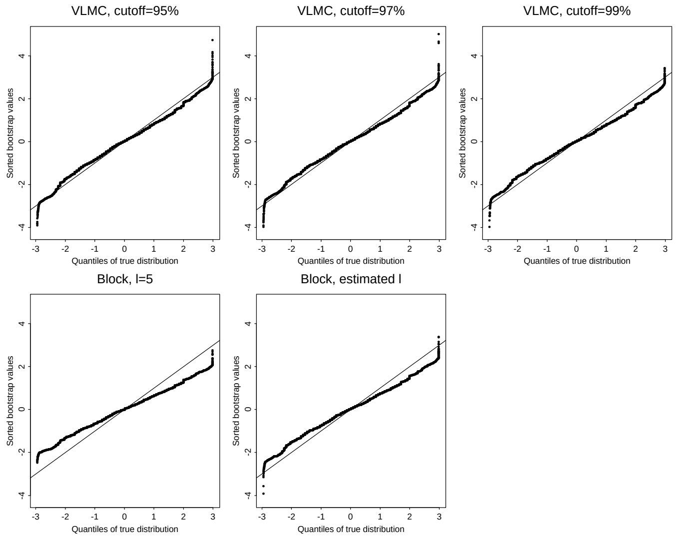

Figure 6.2: Bootstrap distribution estimation of $(\overline{X}_n - \mathbb{E}[X_t])/\sigma_n$ by $(\overline{X}_n^* - \mathbb{E}^*[\overline{X}_n^*])/\sigma_n$ for $n = 128$ $[\sigma_n = (\mathrm{Var}(\overline{X}_n))^{1/2}]$. The target is indicated by the line. VLMC-sieve bootstrap with cutoffs as $\chi^2_{1;\alpha}/2$ with $\alpha = 0.95, 0.97, 0.99$; block bootstrap with $\ell = 5$ and estimated $\hat{\ell}$. 100 simulation runs, 500 bootstrap replicates per simulation run.

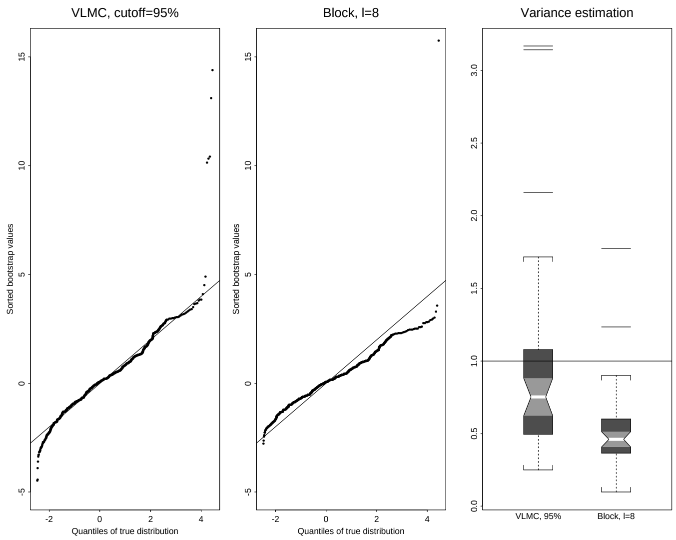

Figure 6.3: Bootstrap distribution and variance estimation of $(\hat{\theta} - \mathbb{E}[\hat{\theta}])/\sigma_n$ by $(\hat{\theta}^* - \mathbb{E}^*[\hat{\theta}^*])/\sigma_n$ for $\hat{\theta}$ from (S2) and $n = 512$ $[\sigma_n = (\mathrm{Var}(\hat{\theta}))^{1/2}]$. Two left panels: QQ-plots with target indicated by the line. Right panel: boxplots with target indicated by the horizontal line. 50 simulation runs, 200 bootstrap replicates per simulation run.

We also examined construction of two-sided confidence interval with the estimator (S1) for $\theta = \mathbb{E}[X_t] = 1/2$ on nominal coverage level 0.9 for sample size $n = 128$. We considered first order accurate block and VLMC-sieve bootstraps and higher order accurate intervals with a version of BC_a for the block bootstrap [see Götze and Künsch, 1996] and the double bootstrap for the VLMC method, analogously to the AR-sieve scheme in section 4.2, formula (4.5) [with the same tuning parameter for the second level bootstrap as in the first level]. The tuning parameters of the methods correspond to the upper left and lower right picture in Figure 6.2. Coverage probabilities of confidence intervals with median and mean absolute deviation of their lengths are given in Table 6.1. For the block bootstrap,

	Block, $\hat{\ell}$	VLMC, 95%	Block BC_a, $\hat{\ell}$	Double VLMC, 95%
coverage	0.71	0.74	0.75	0.84
median(length)	0.262	0.276	0.258	0.368
MAD(length)	0.041	0.063	0.041	0.122

Table 6.1: Coverage probabilities for two-sided confidence interval on nominal 90% level with median and mean absolute deviation (MAD) of their lengths. Sample size $n = 128$. Based on 100 simulations, 500 first level bootstrap replicates; double VLMC bootstrap calibration with 100 first and 100 second level bootstrap replicates.

the asymptotically second order BC_a method increases performance with weak significance compared to the first order block interval. Compared to the first order VLMC and to any of the block methods, the double VLMC bootstrap improves with strong significance upon coverage. On average, it corrects the nominal 90% to the 97.3% coverage level; a calibration for one typical sample is shown in Figure 6.4, yielding the corrected coverage level 96.7%.

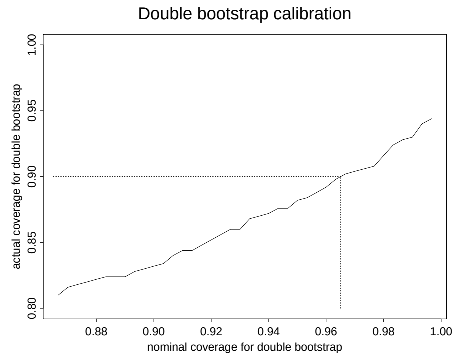

Figure 6.4: Double VLMC bootstrap calibration for one typical sample of size $n = 128$. On the x-axis: nominal coverage level for second level bootstrap based on first level bootstrapped data; on the y-axis, corresponding actual coverage level. [These are the quantities $1 - q$ and $\hat{a}(1 - q)$ in (4.4); the solid and dotted lines indicate the function $\hat{a}(1 - q)$ and the corrected value $\hat{s}_{0.90} = 0.967$, respectively]. Based on 500 first and 500 second level bootstrap replicates.

7 A local bootstrap for conditional mean estimates

We focus on interval estimation with a local bootstrap for the conditional expectation $\theta(x) = \mathbb{E}[X_t|X_{t-1} = x]$ ($x \in \mathbb{R}$) of a stationary real-valued process $(X_t)_{t\in\mathbb{Z}}$. This case, which we choose for reasons of expository simplicity, can be easily extended to the more general parameter $\mathbb{E}[f(X_t)|X_{t-i_1} = x_1, \ldots, X_{t-i_p} = x_p]$ for a chosen set of p lagged indices $t - i_1, \ldots, t - i_p$ and $f : \mathbb{R} \to \mathbb{R}$. Given data $X_1, \ldots, X_n$, consider the kernel estimate

$$\hat{\theta}_h(x) = \frac{\sum_{t=2}^{n} W_{t,h}(x)X_t}{\sum_{t=2}^{n} W_{t,h}(x)},\ W_{t,h}(x) = K(\frac{x - X_{t-1}}{h})$$

with bandwidth h.

For bootstrapping $\hat{\theta}_h(x)$ one can resort to resampling in a local regression framework,

$$X_t^{*L} \sim \hat{F}_{X_{t-1}}\ (t = 2, \ldots, n),\ \text{independently from } X_s^{*L}\ (s \neq t),$$

where $\hat{F}_{x,b}(\cdot) = \sum_{t=2}^{n} W_{t,b}(x)1_{[X_t\leq\cdot]}/\sum_{t=2}^{n} W_{t,b}(x)$ is an estimate of the conditional cumulative distribution of X_t given $X_{t-1} = x$; b is a [pilot-] bandwidth and $W_{.,.}(\cdot)$ as above. Thus, the resampling is driven independently by the estimated $\{\hat{F}_{x,b}; x \in \mathbb{R}\}$ which are allowed to change *locally*.

The bootstrapped kernel estimator $\hat{\theta}_h(x)$ is then given from the regression-type data $(X_1, X_2^{*L}), (X_2, X_3^{*L}), \ldots, (X_{n-1}, X_n^{*L})$,

$$\hat{\theta}_h^{*L}(x) = \frac{\sum_{t=2}^{n} W_{t,h}(x)X_t^{*L}}{\sum_{t=2}^{n} W_{t,h}(x)}.$$

Such an approach was initiated by Neumann and Kreiss (1998) and Paparoditis and Politis (1999b).

A way to see why the local bootstrap works is to consider the asymptotic distribution of the kernel estimator $\hat{\theta}_h(x)$ which depends only on the marginal distribution of X_t, the conditional distribution of X_t given X_{t-1} and the known form of the kernel. The local bootstrap is able to estimate all these features.

This is only asymptotically true and for any finite sample size n, the variance of $\hat{\theta}_h(x)$ depends on the n-dimensional distribution of $(X_t)_{t\in\mathbb{Z}}$, in a specific way. By construction, the local bootstrap is not able to pick-up dependencies beyond the conditional distribution of X_t given X_{t-1}. On the other hand, the block bootstrap estimates consistently the ℓ-dimensional distribution of $(X_t)_{t\in\mathbb{Z}}$ and it is shown in Accola (1998) that it achieves a better rate for estimating $\text{Var}(\hat{\theta}_h(x))$ than the local bootstrap.

Neumann and Kreiss (1998) and Neumann (1998) show the consistency for confidence regions for $\theta(x)$ which are *simultaneous* over x, constructed with a related local bootstrap scheme. The corresponding rate of convergence is $1/\sqrt{nh}$ as for the pointwise case. This is a very important result since analytical simultaneous approximations tend to a limiting extreme value distribution with the extremely slow rate of $1/\log(n)$. Thus, the analytic approach via the limiting distribution is far inferior than the local bootstrap construction.

7.1 Tuning parameter and range of applicability

The tuning parameter of the local bootstrap is the pilot bandwidth b. A simple approach is to choose $b = h$, where h is the pre-chosen bandwidth of the estimator $\hat{\theta}_h(x)$. When b

is chosen of larger order than h, an asymptotically non-negligible bias $\mathbf{E}[\hat{\theta}_h(x)] - \theta(x)$ can be estimated with the local bootstrap, see Paparoditis and Politis (1999b). The role of the pilot-bandwidth is for estimating the conditional distribution of X_t given X_{t-1}: this is relatively easy due to the restriction to a two-dimensional marginal problem, and the procedure is not very sensitive to specification of this pilot-bandwidth. The local bootstrap does not require any tuning parameter for estimating more complicated dependencies. This is an important practical advantage of the method.

The local bootstrap is proven to be consistent whenever $(X_t)_{t\in\mathbb{Z}}$ is a short-range dependent process, see Paparoditis and Politis (1999b) and Ango Nze, Bühlmann and Doukhan (1999).

7.2 Local versus block bootstrap for simulated series

From a practical point of view it is interesting to see whether a bootstrap taking time series effects into account, say the block bootstrap, is advantageous. We consider here a

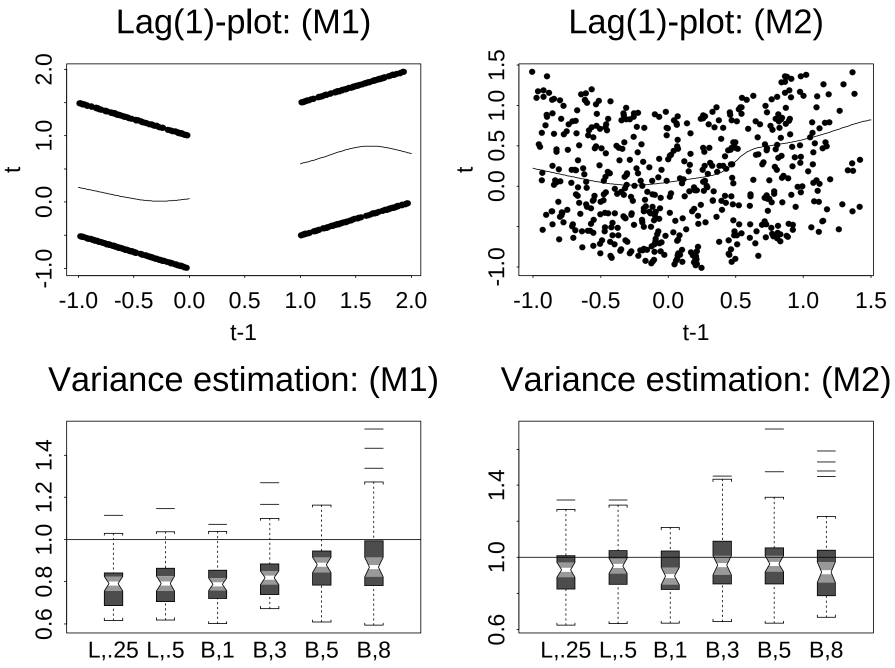

Figure 7.1: Top row: Lag(1) scatter-plot of X_t versus X_{t-1} with the line indicating the curve estimator $\hat{\theta}_h(\cdot)$ with $h = 0.25$. Bottom row: Bootstrap variance estimates $\mathrm{Var}^*(\hat{\theta}_h^*(x))/\mathrm{Var}(\hat{\theta}_h(x))$ at $x = 1.60$ and $x = 0.76$ for (M1) and (M2), respectively [the target is indicated by the horizontal line]; 'L' for local bootstrap with pilot bandwidths 0.25 and 0.5, 'B' for block bootstrap with blocklengths 1, 3, 5 and 8. Everything for sample size $n = 512$.

simulation experiment which deals with a bilinear model,

$$X_t = 0.5\varepsilon_{t-1}X_{t-1} + \varepsilon_t, \tag{7.1}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d. innovation sequence with ε_t independent from $\{X_s; s < t\}$. We consider here the cases, where

(M1) ε_t i.i.d. $\sim$ Rademacher, i.e., $\mathbb{P}[\varepsilon_t = 1] = \mathbb{P}[\varepsilon_t = -1] = 1/2$,

(M2) ε_t i.i.d. $\sim$ Uniform$([-1, 1])$.

Maybe due to the discreteness of the innovations, we found empirically that estimation of $\mathrm{Var}(\hat{\theta}_h(x))$ is harder in (M1) than (M2). Figure 7.1 shows for sample size $n = 512$ the lag(1) scatter-plot of X_t versus X_{t-1} with the estimator $\hat{\theta}_h(\cdot)$ with the standard Gaussian kernel for bandwidth $h = 0.25$. Expressing the bottom row of Figure 7.1 more quantitatively is as follows. In (M1) the block bootstrap with $\ell = 5$ ('B,5') performs best, with respect to MSE: it has about 40% lower MSE than the best local bootstrap with $b = 0.5$ ('L,.5') and the two-sided paired Wilcoxon test yields a p-value of 0.002 for the null-hypothesis of equal MSE for 'L,.5' and 'B,5', in favor of 'B,5'. Comparing any of the local with any of the other block bootstraps with $\ell = 1, 3, 8$ shows no significant difference. In (M2), the block bootstrap with $\ell = 1$ [which is a regression bootstrap under independence] performs best having about 9% lower MSE than the best local bootstrap with $b = 0.5$, but which turns out to be a non-significant difference. Excluding 'B,8' with unreasonably large blocklength [having a significant disadvantage against local bootstraps], any of the local compared with any of the other block bootstraps with $\ell = 1, 3, 5$ shows no significant difference. We conclude for this specific bilinear model that in the easier case (M2), local and block bootstrap are equally good [when excluding the unreasonable blocklength $\ell = 8$]. This is contrasted a bit by the hard case (M1) where the block bootstrap is always as good as local bootstraps and better, would we have known the good blocklength $\ell = 3$.

8 Conclusions

We summarize here our view and empirical results for the block, two types of sieve and a local bootstrap.

The block bootstrap is the most general method. From theory, it is able to cope with very many situations. A further advantage is the simple implementation of resampling, which is no more difficult than in Efron's i.i.d. bootstrap. Disadvantages of the method include the following. The block bootstrap sample should not be viewed as a reasonable copy of a stochastic process: it isn't stationary and exhibits artifacts where resampled blocks are linked together. This implies that the plug-in rule for bootstrapping an estimator $\hat{\theta}$ is not appropriate: the bootstrapped estimator and its computing routine might have to be redesigned, since a pre-vectorization of the data is highly recommended, provided that it is possible or feasible. As a general nonparametric scheme, the block bootstrap can be outperformed in various niches of stationary time series, e.g. for linear time series [see section 4] and for categorical processes [see section 6]. Second order accuracy for a confidence interval has been justified with the approach of studentizing and BC_a correction; particularly the first seems to cause notorious difficulties for finite samples, the latter was found to yield marginal improvement in a simulated example. Double bootstrapping seems not promising since the block bootstrap in the first iteration is not an appropriate time series sample mimicking the underlying data generating process.

Sieve bootstraps in general construct resamples which are sampled from a reasonable time series model. This implies two advantages: the plug-in rule is employed for defining and computing the bootstrapped estimator, and the double bootstrap potentially leads to higher order accuracy. Good sieve bootstraps, as the AR- or VLMC-sieve schemes, are adapting to the degree of dependence: their accuracy improves with decreasing dependence, see formulae (4.3) and (6.2). This is not the case with the block bootstrap, as seen from formula (3.5). Also, sieve bootstraps seem generally less sensitive to selection of a model in the sieve than the block bootstrap to the blocklength.

The AR-sieve bootstrap is clearly best if the data-generating process is a linear time series, representable as an AR(∞) as in (4.1). The method is easy to implement, due to the simplicity of fitting an AR model.

The VLMC-sieve bootstrap is best for categorical processes, particularly when dependence decays exponentially with increasing lags. The disadvantage is the difficulty to construct the resample: the context algorithm which is used for this task is computationally quite efficient needing only $O(n \log(n))$ essential operations, but the algorithm is not easily available yet. Double bootstrapping was successful in a simulated example.

The local bootstrap from section 7 is restricted to nonparametric estimation procedures having slower rate of convergence than $1/\sqrt{n}$. Its advantage is simplicity, since no tuning parameter, governing strength of dependence of the data generating process, has to be specified. On the other hand, this also indicates its weakness and lack of ability to mimic dependence properly. Although designed as a regression bootstrap in the independent set-up, it is consistent and hence robust against some form of dependence. In the latter case, it can be outperformed [with the block bootstrap].

9 Other results and notes to references

We complement our selective exposition by briefly pointing to some additional references. Viewed from a different angle than our review and comparison, Efron and Tibshirani (1993, Chs.8.5–8.6), Shao and Tu (1995, Ch.9), Li and Maddala (1996), Davison and Hinkley (1997, Ch.8), discuss bootstrap methods for dependent data.

Literature about the block bootstrap is quite extensive. A review from the earlier area of the field can be found in Léger, Politis and Romano (1992). Refinement of Künsch's (1989) results, aiming for minimal assumptions, is given in Radulović (1996a). Various results in empirical processes include Bühlmann (1994, 1995), Radulović (1996b, 1998) and Peligrad (1998). Lahiri (1996a) proves second order correctness of the block bootstrap for the case where $\hat{\theta}$ is an M-estimator in a linear regression model with dependent noise. The block bootstrap technique is also applicable for spatial processes, see Politis and Romano (1993). A version of the block bootstrap achieving stationarity for the bootstrap sample, the so-called stationary bootstrap, was given by Politis and Romano (1994). Lahiri (1999) shows rigorously that the block bootstrap is better than the stationary bootstrap. Carlstein, Do, Hall, Hesterberg and Künsch (1998) propose a linking scheme for blocks to be resampled: they argue in case of $\hat{\theta} = \overline{X}_n$, that such a procedure has lower mean square error for variance estimation.

Related to the block bootstrap are subsampling methods. The work by Carlstein (1986) can be viewed as a predecessor of the block bootstrap for variance estimation:

Künsch (1989) argues that in case where the statistic $\hat{\theta}$ is asymptotically normal, the block bootstrap is better than subsampling. In a remarkable paper, Politis and Romano (1994) show that subsampling is much more generally applicable than block bootstraps methods: namely in essentially all cases where $\hat{\theta}$ has some non-degenerated limiting distribution. Again, in case of a normal limit, subsampling is inferior to block bootstrapping. Other results about subsampling can be found in the book by Politis, Romano and Wolf (1999).

Model based bootstrapping with a nonparametric AR(1)-model with heteroscedastic innovations is discussed in Franke, Kreiss and Mammen (1997) and Franke, Kreiss, Mammen and Neumann (1998). The latter discusses that such a bootstrap can be used for accurate construction of simultaneous confidence bands of the autoregression function $m(x) = \mathbb{E}[X_t|X_{t-1} = x]$; note that the same can be achieved [in first order] by a local bootstrap as in section 7. Bootstrapping based on a finite order Markov model, similarly as Rajarshi's (1990) scheme, is also studied by Paparoditis and Politis (1997).

For the AR-sieve bootstrap, empirical process results are given in Bickel and Bühlmann (1999) by establishing a weak notion of mixing for the bootstrapped process. The nonstationary case where $X_t = m_t + Z_t$ $(t \in \mathbb{Z})$ with $(m_t)_{t\in\mathbb{Z}}$ a slowly varying deterministic trend and $(Z_t)_{t\in\mathbb{Z}}$ an AR(∞) noise-process is studied in Bühlmann (1998) with an AR-sieve bootstrap for the trend estimate.

Combining model or sieve based methods with the block bootstrap was suggested, and named as 'post-blackening', by Davison and Hinkley (1997, Ch.8.2). The idea is to pre-whiten the time series with a model or sieve based approach and then apply the block bootstrap to the hopefully less dependent, whitened residuals: block resampling of these residuals and inverting the whitening operation then yields the post-blackened resample mimicking the original observations.

Another way of bootstrapping stationary linear time series was proposed by Dahlhaus and Janas (1996): they resample independently periodogram values in the frequency domain according to a spectral density estimate. By construction, the resampling only considers the autocovariance structure and is thus restricted to linear time series. The idea of independent resampling in the frequency domain already appeared in Franke and Härdle (1992) for bootstrapping a spectral density estimator; a modification thereof with a bootstrap scheme of local type was given by Paparoditis and Politis (1999a).

Breiman's (1996) strategy of bagging ['bootstrap aggregating'] an unstable predictor based on independent observations has been found empirically successful when applied to a time series version of MARS for prediction, see Ray and Chen (1998) : they apply a local bootstrap version, similarly as in section 7.

References

[1] Accola, C. (1998). Bootstrap für die bedingte Erwartung bei Zeitreihen. (In German). Diploma thesis, ETH Zürich.

[2] Ango Nze, P., Bühlmann, P. and Doukhan, P. (1999). Nonparametric regression estimation beyond mixing and association. Preprint, ETH Zürich.

[3] Beran, R. (1987). Prepivoting to reduce level error of confidence sets. Biometrika **74**, 457–468.

[4] Bickel, P.J. and Bühlmann, P. (1996). What is a linear process? Proc. Nat. Academy of Sciences U.S.A. **93**, 12128–12131.

[5] Bickel, P.J. and Bühlmann, P. (1997). Closure of linear processes. J. Theoret. Probab. **10**, 445–479.

[6] Bickel, P.J. and Bühlmann, P. (1999). A new mixing notion and functional central limit theorems for a sieve bootstrap in time series. Bernoulli **5**, 413–446.

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

[8] Breiman, L. (1996). Bagging predictors. Machine Learning **26**, 123–140.

[9] Brillinger, D.R. (1997). Random process methods and environmental data: the 1996 Hunter Lecture. Environmetrics **8**, 269–281.

[10] Bühlmann, P. (1994). Blockwise bootstrapped empirical process for stationary sequences. Ann. Statist. **22**, 995–1012.

[11] Bühlmann, P. (1995) The blockwise bootstrap for general empirical processes of stationary sequences. Stoch. Process. Appl. **58**, 247–265.

[12] Bühlmann, P. (1997). Sieve bootstrap for time series. Bernoulli **3**, 123–148.

[13] Bühlmann, P. (1998). Sieve bootstrap for smoothing in non-stationary time series. Ann. Statist. **26**, 48–83.

[14] Bühlmann, P. (1999). Model selection for variable length Markov chains and tuning the context algorithm. To appear in Ann. Inst. Statist. Math.

[15] Bühlmann, P. and Künsch, H.R. (1995). The blockwise bootstrap for general parameters of a stationary time series. Scand. J. Statist. **22**, 35–54.

[16] Bühlmann, P. and Künsch, H.R. (1999). Block length selection in the bootstrap for time series. To appear in Comput. Statist. Data Anal.

[17] Bühlmann, P. and Wyner, A.J. (1999). Variable length Markov chains. Ann. Statist. **27**, No.2.

[18] 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.

[19] Carlstein, E., Do, K.-A., Hall, P., Hesterberg, T. and Künsch, H.R. (1998). Matched-block bootstrap for dependent data. Bernoulli **4**, 305–328.

[20] Choi, E. and Hall, P. (1999). Bootstrap confidence regions computed from autoregressions of arbitrary order. Preprint, Australian National University.

[21] Dahlhaus, R. and Janas, D. (1996). A frequency domain bootstrap for ratio statistics in time series analysis. Ann. Statist. **24**, 1934–1963.

[22] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications.* Cambridge University Press.

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

[24] Franke, J. Kreiss, J.-P. and Mammen, E. (1997). Bootstrap of kernel smoothing in nonlinear time series. Preprint, Universität Kaiserslautern.

[25] Franke, J. Kreiss, J.-P., Mammen, E. and Neumann, M.H. (1998). Properties of the nonparametric autoregressive bootstrap. Preprint, Universität Kaiserslautern.

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

[27] Efron, B. (1987). Better bootstrap confidence intervals. J. Amer. Statist. Assoc. **82**, 171–185.

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

[29] Ferrari, F. (1999). Estimation of general stationary processes by variable length Markov chains. Preprint. ETH Zürich.

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

[31] Götze, F. and Künsch, H.R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. Ann. Statist. **24**, 1914–1933.

[32] Grenander, U. (1981). *Abstract Inference.* Wiley, New York.

[33] Hall, P. (1985). Resampling a coverage pattern. Stoch. Process. Appl. **20**, 231–246.

[34] Hall, P. (1986). On the bootstrap and confidence intervals. Ann. Statist. **14**, 1431–1452.

[35] Hall, P., Horowitz, J.L. and Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. Biometrika **82**, 561–574.

[36] Hall, P., Jing, B.-Y. and Lahiri, S.N. (1998). On the sampling window method under long range dependence. Statist. Sinica **8**, 1189–1204.

[37] Hart, J.D. (1995). Some automated methods of smoothing time-dependent data. J. Nonpar. Statist. **6**, 115–142.

[38] Kreiss, J.-P. (1992). Bootstrap procedures for AR(∞)-processes. In *Bootstrapping and Related Techniques,* Eds. Jöckel, K.-H., Rothe, G. and Sendler, W., pp. 107–113. Lect. Notes in Economics and Math. Systems **376**. Springer, Heidelberg.

[39] Kreiss, J.-P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models. J. Time Ser. Anal. **13**, 297–317.

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

[41] Lahiri, S.N. (1993). On the moving block bootstrap under long range dependence. Statist. Probab. Lett. **18**, 405–413.

[42] Lahiri, S.N. (1995). On the asymptotic behaviour of the moving block bootstrap for normalized sums of heavy-tail random variables. Ann. Statist. **23**, 1331–1349.

[43] Lahiri, S.N. (1996). On Edgeworth expansion and moving block bootstrap for studentized M-estimators in multiple linear regression models. J. Multivar. Anal. **56**, 42–59.

[44] Lahiri, S.N. (1996b). On empirical choice of the optimal block size for block bootstrap methods. Preprint. University of Iowa, Ames.

[45] Lahiri, S.N. (1999). Theoretical comparisons of block bootstrap methods. Ann. Statist. **27**, 386–404.

[46] Léger, C., Politis, D.N. and Romano, J.P. (1992). Bootstrap technology and applications. Technometrics **34**, 378–398.

[47] Li, Hongyi and Maddala, G.S. (1996). Bootstrapping Time Series Models. Economet. Rev. **15**, 115–158.

[48] Loh, W. (1987). Calibrating confidence coefficients. J. Amer. Statist. Assoc. **82**, 155–162.

[49] Neumann, M.H. (1998). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. Ann. Statist. **26**, 2014–2048.

[50] Neumann, M.H., Kreiss, J.-P. (1998). Regression-type inference in nonparametric autoregression. Ann. Statist. **26**, 1570–1613.

[51] Paparoditis, E. and Politis, D.N. (1997). The local bootstrap for Markov processes. Preprint, University of Cyprus.

[52] Paparoditis, E. and Politis, D.N. (1999a). The local bootstrap for periodogram statistics. J. Time Ser. Anal. **20**, 193–222.

[53] Paparoditis, E. and Politis, D.N. (1999b). The local bootstrap for kernel estimators under general dependence conditions. To appear in Ann. Inst. Statist. Math.

[54] Peligrad, M. (1998). On the blockwise bootstrap for empirical processes for stationary sequences. Ann. Probab. **26**, 877–901.

[55] Politis, D.N. and Romano, J.P. (1992). A general resampling scheme for triangular arrays of α-mixing random variables. Ann. Statist. **20**, 1985–2007.

[56] Politis, D.N. and Romano, J.P. (1993). Nonparametric resampling for homogeneous strong mixing random fields. J. Multivar. Anal. **47**, 301–328.

[57] Politis, D.N. and Romano, J.P. (1994). The stationary bootstrap. J. Amer. Statist. Assoc. **89**, 1303–1313.

[58] Politis, D.N. and Romano, J.P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. Ann. Statist. **22**, 2031–2050.

[59] Politis, D.N., Romano, J.P. and Wolf, M. (1999). *Subsampling.* To appear in Springer.

[60] Radulović, D. (1996a). The bootstrap of the mean for strong mixing sequences under minimal conditions. Statist. Prob. Lett. **28**, 65–72.

[61] Radulović, D. (1996b). The bootstrap for empirical processes based on stationary observations. Stoch. Process. Appl. **65**, 259–279.

[62] Radulović, D. (1998). The bootstrap of empirical processes for α-mixing sequences. In *High dimensional probability,* Eds. Eberlein, E., Hahn, M. and Talagrand, M., pp. 315–330. Progress in Probability, Vol. 43. Birkhäuser, Basel.

[63] Ray, B.K. and Chen, L. (1998). Bootstrapping ASTAR models. Preprint. New Jersey Institute of Technology.

[64] Rissanen, J. (1983). A universal data compression system. IEEE Trans. Inform. Theory **IT-29**, 656–664.

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

[66] Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Ann. Statist. **8**, 147–164.