Statistical Science
2002, Vol. 17, No. 1, 52–72

Bootstraps for Time Series

Peter Bühlmann

Abstract. We review and compare block, sieve and local bootstraps for time series and thereby illuminate theoretical aspects of the procedures as well as their performance on finite-sample data. Our view is *selective* with the intention of providing a new and fair picture of some particular aspects of bootstrapping time series.

The generality of the block bootstrap is contrasted with sieve bootstraps. We discuss implementational advantages and disadvantages. We argue that two types of sieve often outperform the block method, each of them in its own important niche, namely linear and categorical processes. 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, Studentizing.

1. INTRODUCTION

Bootstrapping can be viewed as simulating a statistic or statistical procedure from an estimated distribution $\hat{P}_n$ of observed data $X_1, \dots, X_n$. Under dependence, the construction of $\hat{P}_n$ is more complicated and far less obvious than in Efron's (1979) seminal proposal for the independent setup. We discuss here mainly block, sieve and local bootstraps, which are all in a certain sense nonparametric and model-free. The purpose is to obtain a fair picture of strengths and weaknesses of such different time series bootstraps. To do so, we focus on theoretical aspects as well as on performance for finite sample data. So far, very little attention has been paid to an overall perspective when comparing different schemes. In that respect, our *selective* view offers valuable new insights and makes our comparative exposition rather different from those of Léger, Politis and Romano (1992), Efron and Tibshirani (1993, Chapters 8.5–8.6), Shao and Tu (1995, Chapter 9), Li and Maddala (1996) or Davison and Hinkley (1997, Chapter 8).

Extracting information from data is formalized here with a scalar-, vector- or curve-valued estimator $\hat{\theta}$. Estimation of the sampling distribution of $\hat{\theta}$, or pivotized or Studentized versions thereof, is essential for statistical inference. With time series data, this task is much more difficult than for independent observations and methods based on analytic derivations quickly become complicated. For example, consider an estimator $\hat{\theta}$ which is asymptotically normally distributed around a finite-dimensional parameter θ of interest: under suitable conditions and assuming stationarity of the data $X_1, \dots, X_n$,

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

In contrast to the i.i.d. setup, the asymptotic variance σ_∞^2 is an infinite-dimensional object, involving an infinite sum of covariances, which is generally not estimable with convergence rate $1/\sqrt{n}$. As a simple example,

$$\text{if } \hat{\theta} = n^{-1} \sum_{t=1}^{n} X_t, \quad \sigma_\infty^2 = \sum_{k=-\infty}^{\infty} \operatorname{Cov}(X_0, X_k).$$

The asymptotic variance is thus the spectral density of the data-generating process at zero (normalized by the factor 2π). As another example,

$$\text{if } \hat{\theta} = \operatorname{med}(X_1, \dots, X_n),$$

$$\sigma_\infty^2 = \sum_{k=-\infty}^{\infty} \operatorname{Cov}(\operatorname{IF}(X_0), \operatorname{IF}(X_k)),$$

$$\operatorname{IF}(x) = \frac{\operatorname{sign}(x - \theta)}{2f(\theta)},$$

Peter Bühlmann is Associate Professor at Seminar für Statistik, ETH Zürich, CH-8092 Zürich, Switzerland (e-mail: buhlmann@stat.math.ethz.ch).

where $\theta = F^{-1}(1/2)$ is the median of the cumulative marginal distribution F of X_t having density f. Here, the spectral density of the process with influence functions $(\mathrm{IF}(X_t))_{t\in\mathbb{Z}}$ is involved, that is, an instantaneous unknown transform of the process $(X_t)_{t\in\mathbb{Z}}$. It would be very awkward to estimate the unknown density f and θ to get an estimate of $\mathrm{IF}(\cdot)$ and then of its spectral density. Bootstraps have the advantage of consistently estimating the asymptotic variance and distribution of $\sqrt{n}(\hat{\theta} - \theta)$ *automatically*.

Consistency, or first-order accuracy, is defined by requiring consistent estimation of the limiting distribution of $\hat{\theta}$. More precisely, for an $\mathbb{R}^q$-valued estimator $\hat{\theta}$,

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

where $a_n(\hat{\theta} - \theta)$ converges to a nondegenerate limiting distribution. Here, the symbol "$\le$" is defined componentwise and, as usual, the asterisk * denotes a bootstrap quantity. The centering value θ^*, which is a constant conditional on the original observations $X_1, \ldots, X_n$, is typically *not* chosen to be $\hat{\theta}$ as in Efron's i.i.d. bootstrap; details are given later when specifying particular time series bootstraps.

For example, if $\hat{\theta}$ is the sample mean or median, the limiting distribution is $\mathcal{N}(0, \sigma_\infty^2)$ as in (1.1) with σ_∞^2 from above (provided that some regularity conditions hold). Consistency is then implied by

$$\sqrt{n}(\hat{\theta}^* - \theta^*) \Rightarrow \mathcal{N}(0, \sigma_\infty^2) \quad \text{in probability, as } n \to \infty,$$

saying that the limiting distributions of the bootstrapped and original estimator coincide. This convergence typically requires, among other things, that the bootstrap variance is asymptotically correct,

$$n\mathrm{Var}^*(\hat{\theta}^*) = \sigma_\infty^2 + o_P(1), \quad n \to \infty.$$

Since often $\sigma_\infty^2 = \lim_n n\mathrm{Var}(\hat{\theta})$, this can be viewed as a convergence of standardized variances,

$$n\mathrm{Var}^*(\hat{\theta}^*) - n\mathrm{Var}(\hat{\theta}) = o_P(1), \quad n \to \infty.$$

Bootstrap consistency in (1.2) usually holds when $\hat{\theta}$ is asymptotically normal. As much as approximating the distribution of an estimator $\hat{\theta}$, the bootstrap procedure also allows $\mathrm{Var}(\hat{\theta})$ to be approximated by the bootstrap variance $\mathrm{Var}^*(\hat{\theta}^*)$. The accuracy for distribution estimation in (1.2) is driven by the accuracy of the bootstrap variance

$$(1.3)\qquad a_n^2\mathrm{Var}^*(\hat{\theta}^*) - a_n^2\mathrm{Var}(\hat{\theta}),$$

provided that Edgeworth expansions for $\hat{\theta}$ and $\hat{\theta}^*$ are valid. Usually, the infinite-dimensional character of the limiting variance makes this problem of accurate bootstrap variance estimation much harder than in the independent setting.

Of course, as in the case with independent data, time series bootstraps also offer the advantage of higher order accuracy improving upon estimated normal approximations as in (1.1). The approximation is then estimated for Studentized versions of $\hat{\theta}$, or a confidence interval is adjusted with the BC_a [bias corrected and accelerated (Efron, 1987)] or a double bootstrap correction. However, for *finite* sample situations, first-order schemes often may be as accurate as their second-order counterparts, and a good bound in (1.3) is then desired. A substantial part of the paper is devoted to the discussion of first-order accuracy, but we also include aspects of second-order correctness.

2. 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 to preserve the original time series structure within a block. Such an idea appears in Hall (1985), but the breakthrough of the block bootstrap is given by Künsch's (1989) paper, explaining in detail how and why such a bootstrap works.

2.1 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 parameter θ which is a functional of 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_0, X_1), corresponding to $m = 2$. Consider then vectors of consecutive observations

$$(2.1)\quad Y_t = (X_{t-m+1}, \ldots, X_t), \qquad t = m, \ldots, n,$$

and construct the block-resampling on the basis of these vectorized observations as follows. 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 the number of blocks $n - m + 1 = k\ell$ with $k \in \mathbb{N}$. Then, resample k blocks independently with replacement,

$$(2.2)\quad \begin{aligned} &Y_{S_1+1}, \ldots, Y_{S_1+\ell}, \\ &Y_{S_2+1}, \ldots, Y_{S_2+\ell}, \ldots, Y_{S_k+1}, \ldots, Y_{S_k+\ell}, \end{aligned}$$

where the block-starting points $S_1, \dots, S_k$ are i.i.d. Uniform($\{m-1, \dots, n-\ell\}$) on the possible starting locations. If the number of blocks $n-m+1$ is not a multiple of ℓ, we resample $k = \lfloor (n-m+1)/\ell \rfloor + 1$ blocks but use only a portion of the kth block to get $n-m+1$ resampled m-vectors in total. These resampled blocks of m-vectors in (2.2) could be referred to the block bootstrap sample. However, as we will see, the block bootstrapped estimator is not defined by a plug-in rule and the notion of a bootstrap sample is then not so clear.

A good definition of the block bootstrapped estimator is not entirely straightforward. The vectorization in (2.1) is typically associated with the estimator so that

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

For example, it is often possible to represent the estimator as

$$\hat\theta = T(F_n^{(m)}), \tag{2.3}$$

where $F_n^{(m)}(\cdot) = (n-m+1)^{-1}\sum_{t=m}^n \mathbb{1}_{[Y_t \le \cdot]}$ 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 $\theta = \mathrm{Corr}(X_0, X_1)$, consider the estimator $\hat\theta = \hat R(1)/\hat R(0)$ with $\hat R(k) = (n-1)^{-1}\sum_{t=1}^{n-1}(X_t - \hat\mu_X)(X_{t+k} - \hat\mu_X)$ ($k \in \{0,1\}$), $\hat\mu_X = (n-1)^{-1}\sum_{t=1}^{n-1} X_t$. This estimator $\hat\theta$ is symmetric in $Y_2, \dots, Y_n$ with $Y_t = (X_{t-1}, X_t)$ and it is of the form (2.3) with $m = 2$. Note that the usual estimator is $\tilde\theta = \tilde R(0)/\tilde R(1)$ with $\tilde R(k) = n^{-1}\sum_{t=1}^n (X_t - \overline{X}_n)(X_{t+k} - \overline{X}_n)$, which is approximately equal to $\hat\theta$, modulo "edge effects."

EXAMPLE B. The GM-estimators (generalized M-estimators) in an AR(p) model can be written in the form (2.3) with $m = p+1$. They are defined implicitly, analogously to the normal equations, by

$$\sum_{t=p+1}^n w_t \psi\big((X_t - \hat\phi_1 X_{t-1} - \cdots - \hat\phi_p X_{t-p})\sigma^{-1}\big)$$
$$\times (X_{t-1}, \dots, X_{t-p})^T = 0,$$

where ψ: $\mathbb{R} \to \mathbb{R}$, σ^2 is the innovation variance and $(w_t)_{t=p+1}^n$ is a sequence of appropriate weights. See Martin and Yohai (1986). Besides the Gaussian maximum likelihood estimator (MLE) with $\psi(x) = x$, this includes estimators being robust against innovation and lagged-value outliers.

The block bootstrapped estimator corresponding to (2.3) is defined as

$$\hat\theta^{*B} = T(F_n^{(m)*B}),$$
$$F_n^{(m)*B}(\cdot) = (n-m+1)^{-1}\sum_{i=1}^k \sum_{t=S_i+1}^{S_i+\ell} \mathbb{1}_{[Y_t \le \cdot]} \tag{2.4}$$

with k and S_i as in (2.2). The centering value θ^* for the block bootstrap in (1.2) is often $\mathbb{E}^{*B}[\hat\theta^{*B}]$, which is generally different from $\hat\theta$. This definition of the block bootstrapped estimator, given by Künsch (1989), can be interpreted as follows. If $\hat\theta = g_{n-m+1}(Y_m, \dots, Y_n)$ is a symmetric function $g_{n-m+1}(\cdot)$ of $n-m+1$ vectorized observations, then

$$\hat\theta^{*B} = g_{n-m+1}(Y_{S_1+1}, \dots, Y_{S_1+\ell}, Y_{S_2+1}, \dots,$$
$$Y_{S_2+\ell}, \dots, Y_{S_k+1}, \dots, Y_{S_k+\ell}),$$

employing a plug-in rule based on the vectorized observations. In particular, the block bootstrapped estimator is defined with Y-variables occurring only in the set of the original vectorized observations. This would *not* be the case without the vectorization step in (2.1). Figure 1 illustrates the artifact of the naive block bootstrap using $m = 1$ instead of the correct $m = 2$ in Example A. The most striking defect of the naive block bootstrap sample are the newly created points in the scatter plot within the rectangles in the

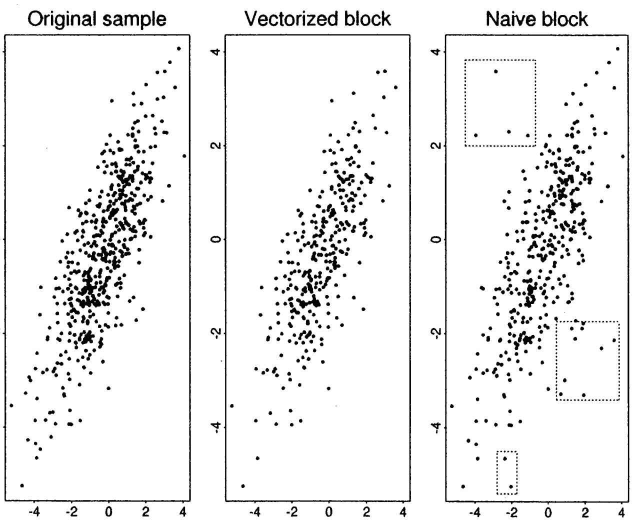

FIG. 1. *Lag*(1) *scatter plots of* (*re*)*samples of size* $n = 512$: (*left panel*) *original sample* (X_{t-1}, X_t), $t = 2, \dots, n$; (*middle panel*) *block bootstrap sample* Y_{S_i+j}, $i = 1, \dots, k = 64$, $j = 1, \dots, \ell = 8$, *from* (2.2) *with* $m = 2$; (*right panel*) *naive block bootstrap sample* $(X_{t-1}^{*nB}, X_t^{*nB})$, $t = 2, \dots, n$, *where* X_t^{*nB} *is the sequentially* t*th value in* (2.2) *with* $m = 1$, $k = 64$, $\ell = 8$; *the points within the rectangles* (*and others*) *do not occur in the plot of the left panel.*

upper left and the lower right corners. A naively block bootstrapped estimator (e.g., for the autocorrelation in Example A) which uses the plug-in rule in conjunction with the naive block bootstrap sample may be strongly affected by these newly created and bad points. As mentioned already, this artifact is not present with the block bootstrap definition in (2.4) based on the vectorized observations.

For the block bootstrap procedure, at least two difficulties remain to be answered case by case:

1. Due to the lack of the plug-in principle, redesigning the *computation* of $\hat{\theta}^{*B}$ is often necessary and can become very inconvenient.
2. Vectorization as in (2.1) is not always appropriate. For example, the MA-parameter in an MA(1) model or the spectral density of a stationary process depends on the entire distribution of the process, corresponding to $m = \infty$.

Whenever problem (1) or (2) becomes too awkward, an ad-hoc solution is to ignore the vectorization step in (2.1) and work with the naive block bootstrap (using $m = 1$). As a result, a substantial efficiency loss of the method may occur. 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).

2.2 Range of Applicability and Accuracy

The block bootstrap is designed to work for general stationary data generating processes $(X_t)_{t\in\mathbb{Z}}$ with $X_t \in \mathbb{R}^d$ $(d \geq 1)$ or taking values in a categorical space. From an asymptotic point of view, the blocklength ℓ should grow, but not too fast, as $n \to \infty$. When restricting to short-range dependent processes (e.g., summable autocovariances or mixing coefficients), the block bootstrap has been theoretically justified in many circumstances: for example, for estimators as in (2.3) with smooth T, cf. Künsch (1989) and Bühlmann (1994). Other references are given in Section 8. Under long-range dependence, some theory and modifications are worked out for the 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 but the bootstrapped statistic has to be corrected with a factor depending on the typically unknown rate of convergence, for example, on the self-similarity parameter in self-similar processes. If $\overline{X}_n$ has a nonnormal limit due to long-range dependence, Hall, Jing and Lahiri (1998) show consistency of a modified block-subsampling procedure. In the case where the observations have a heavy tailed marginal distribution, Lahiri (1995) shows that block bootstrapping with resampling size $m \ll n$ is consistent for the case with $\hat{\theta} = \overline{X}_n$.

Regarding accuracy of the block bootstrap, consider first estimation of the asymptotic variance of $\hat{\theta}$. Künsch (1989) showed that for the mean squared error

$$(2.5)\quad \mathbb{E}[(n\mathrm{Var}^{*B}(\hat{\theta}^{*B}) - n\mathrm{Var}(\hat{\theta}))^2] \sim \mathrm{const}\cdot n^{-2/3},$$

achieved with the rate-optimal blocklength $\ell = \mathrm{const}\cdot n^{1/3}$. Note that this corresponds to (1.3) with $a_n = \sqrt{n}$. The essential assumptions for this result require that $n\mathrm{Var}(\hat{\theta})$ converge to a nondegenerate limiting variance, T in (2.3) be sufficiently smooth and some mixing conditions for the stationary data-generating process $(X_t)_{t\in\mathbb{Z}}$ hold. A bit surprisingly, the rate $n^{-2/3}$ does *not* depend on the "degree of dependence," for example, 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 for this nonadaptivity was already given by Künsch (1989): the block bootstrap variance estimate is asymptotically equivalent to a lag-window spectral density estimator at zero with triangular window,

$$(2.6)\quad n\mathrm{Var}^{*B}(\hat{\theta}^{*B}) \approx \sum_{k=-\ell}^{\ell}\left(1 - \frac{|k|}{\ell}\right)\hat{R}_{\mathrm{IF}}(k),$$

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 $F^{(m)}$; the influence function $\mathrm{IF}(\cdot; F^{(m)})$ is the transformation which asymptotically linearizes a suitably regular estimator [see (2.7)]. However, the triangular form of the window $1 - |k|/\ell$ $(k = -\ell, \dots, 0, \dots, \ell)$ makes it impossible to improve upon the $n^{-2/3}$ MSE-rate. Tapered block bootstraps overcome this limitation; see Künsch's [(1989), formula (2.12)] brief remark and Paparoditis and Politis (2001) for a different, rigorously analyzed proposal.

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 $\ell = \mathrm{const}\cdot n^{1/3}$. [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.] As in variance estimation, the rate of accuracy cannot be improved for time series having geometrically fast decaying dependence properties. Götze and Künsch (1996) also justify a modification of Efron's (1987) BC_a proposal. For finite samples, a second-order accurate method may not always be beneficial. Unfortunately, there is no easy way to judge from data whether a second-order technique pays off. Double block bootstrapping for correcting a first-order bootstrap confidence region is not straightforward because dependence is corrupted at places where blocks join (cf. Davison and Hall, 1993, and Choi and Hall, 2000).

2.3 Choosing a Blocklength ℓ

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

Consider first block bootstrap variance estimation for an estimator $\hat{\theta}$ of the form (2.3). Then

$$\hat{\theta} \approx (n-m+1)^{-1} \sum_{t=m}^{n} \mathrm{IF}(Y_t; F^{(m)}), \qquad (2.7)$$

where $\mathrm{IF}(\cdot;\, F^{(m)})$ is the influence function of $\hat{\theta}$ at $F^{(m)}$. Based on this linearization, formula (2.6) can be shown and rewritten as

$$n\mathrm{Var}^{*B}(\hat{\theta}^{*B}) \approx 2\pi \hat{f}_{\mathrm{IF}}(0), \qquad (2.8)$$

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

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

Bühlmann and Künsch (1999) propose estimation of ℓ_{opt} (or the constant in the expression above) by an iterative plug-in scheme for optimal local bandwidth choice in spectral density estimation at frequency zero, using the asymptotic equivalence in (2.8).

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 estimated optimal 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, the performance on a subsample can be very poor; this inefficiency is demonstrated in a similar context in Section 4.3.

Regarding block bootstrap bias estimation, Lahiri (1999) shows that the asymptotic MSE-optimal blocklengths for bias and variance estimation are the same: estimated blocklengths for variance can thus be used for bias estimation as well.

Automatic choice of the blocklength is at least as difficult as selection of a *local* bandwidth-type tuning parameter in the context of time series. Even worse, (2.8) describes an equivalence to a bandwidth selection problem only asymptotically: the linearization in (2.7) can have a substantial effect for finite sample size. Furthermore, the blocklength ℓ has no practically relevant interpretation and diagnostic tools for it are so far undeveloped.

3. AR-SIEVE BOOTSTRAP FOR STATIONARY LINEAR TIME SERIES

Generally, sieve bootstraps rely on the idea of sieve approximation (Grenander, 1981) for the data-generating process $(X_t)_{t\in\mathbb{Z}}$ by a family of (semi)parametric models. The bootstrap is then nothing other than simulating from a sieve-estimated process.

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

$$X_t - \mu_X = \sum_{j=1}^{\infty} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}, \qquad (3.1)$$

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 $\{X_s; s < t\}$. This is well defined if, for example, $\mathbb{E}[\varepsilon_t^2] < \infty$ and $\sum_{j=1}^{\infty} \phi_j^2 < \infty$.

3.1 The AR-Sieve Bootstrap Procedure

The AR-sieve approximation is constructed with AR(p) models

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

where μ_X and ε_t are as in (3.1). Given data, we first choose an autoregressive order $\hat{p}$, for example, with the Akaike information criterion (AIC) for Gaussian innovations (cf. Shibata, 1980). The remaining parameter of interest $\eta_{\hat{p}} = (\mu_X, (\phi_1, \dots, \phi_{\hat{p}}), F_\varepsilon)$ is semiparametric. Here, F_ε denotes the distribution of the i.i.d. innovations ε_t. The estimates are chosen as follows:

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

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

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

$$R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j},$$

with $\overline{R}_.$ the mean of the available residuals R_t ($t = \hat{p}+1, \dots, n$).

The estimates $\hat{p}$, $\hat{\eta}_{\hat{p}}$ characterize a distribution $\hat{P}_{n;AR}$ for an autoregressive process. It can be represented by the following AR($\hat{p}$) equation:

$$\begin{aligned} &X_t^{*\text{AR-S}} - \hat{\mu}_X \\ (3.2)\quad &= \sum_{j=1}^{\hat{p}} \hat{\phi}_j (X_{t-j}^{*\text{AR-S}} - \hat{\mu}_X) + \varepsilon_t^*, \quad t \in \mathbb{Z}, \end{aligned}$$

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 $X_1^*, \dots, X_n^*$ from the process in (3.2) having distribution $\hat{P}_{n;\text{AR}}$. The computation in practice is as follows. Start with $(X_{-u}^*, \dots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \dots, \hat{\mu}_X)$ with u large, for example, $u = 1{,}000$. Then simulate X_t^* for $t = -u + \hat{p}, \dots, 0, 1, \dots, n$ according to (3.2). Since the estimated process in (3.2) is (with high probability) Markovian and geometrically mixing, the values $X_1^*, \dots, X_n^*$ from our simulated sample are a very good approximation for a sample of the *stationary* distribution of the process in (3.2). The AR-sieve bootstrapped estimator $\hat{\theta}^{*\text{AR-S}}$ is constructed with the plug-in rule. Writing $\hat{\theta} = h_n(X_1, \dots, X_n)$ as a function of the original data $X_1, \dots, X_n$, we define

$$(3.3)\quad \hat{\theta}^{*\text{AR-S}} = h_n(X_1^{*\text{AR-S}}, \dots, X_n^{*\text{AR-S}}).$$

Such a bootstrap was introduced by Kreiss (1992) and further analyzed by Bühlmann (1997), Bickel and Bühlmann (1999) and Choi and Hall (2000).

The centering value θ^* in (1.2) for the AR-sieve bootstrap is obtained as follows. The parameter of interest θ is a functional of the true underlying process $(X_t)_{t\in\mathbb{Z}} \sim P$: $\theta^{*\text{AR-S}}$ is then the same functional evaluated at the estimated $\hat{P}_{n;\text{AR}}$ which generates the bootstrapped process in (3.2).

EXAMPLE A (Continued). For the lag(1)-correlation estimator, $\theta^{*\text{AR-S}} = \text{Corr}^{*\text{AR-S}}(X_t^{*\text{AR-S}}, X_{t+1}^{*\text{AR-S}})$.

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

1. Generate one very long realization $X_1^{*\text{AR-S}}, \dots, X_v^{*\text{AR-S}}$ with $v \gg n$.
2. Use $\hat{\theta}_v^{*\text{AR-S}} = h_v(X_1^{*\text{AR-S}}, \dots, X_v^{*\text{AR-S}})$ as a Monte Carlo approximation of $\theta^{*\text{AR-S}}$.

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

3.2 Range of Applicability and Accuracy

The AR-sieve bootstrap relies heavily on the crucial assumption that the data $X_1, \dots, X_n$ is a finite realization of an AR(∞)-process as in (3.1). In such a setting, consistency as in (1.2) for $\hat{\theta}$ a smooth function of means is given in Bühlmann (1997); the result is extended in Bickel and Bühlmann (1999) for $\hat{\theta}$ as in (2.3). Thereby, the approximating autoregressive order should grow asymptotically, but not too fast, as $n \to \infty$. The AR(∞) representation includes the important class of ARMA models

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

with invertible generating MA-polynomial; that is, $\Psi(z) = 1 + \sum_{k=1}^{q} \psi_k z^k$, $z \in \mathbb{C}$, has its roots outside the unit disk $\{z \in \mathbb{C};\ |z| \le 1\}$. Here $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d. innovation sequence and a few additional regularity conditions, standard in ARMA model theory, have to be made. Of course, there are also many processes which are not representable as an AR(∞): for example, the nonlinear AR(2) in (4.4) or the bilinear model in (6.1) below. Unfortunately, testing for linearity or AR(∞) representation is very delicate: Bickel and Bühlmann (1997) show that the closure of linear or AR-processes is surprisingly large. It reflects the difficulty of judging, for a particular data set, whether the AR-sieve bootstrap will be suitable.

Within the class of linear invertible time series as defined in (3.1), 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 2. 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 $\hat{p}$ from the AIC,

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

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

$$\begin{aligned} &n\mathrm{Var}^{*\text{AR-S}}(\overline{X}_n^{*\text{AR-S}}) - n\mathrm{Var}(\overline{X}_n) \\ &\quad = O_P(n^{-1/2+\kappa}) \quad \text{for any } \kappa > 0. \end{aligned} \tag{3.4}$$

The two results show that the method *adapts automatically* to the decay of the underlying dependence structure, a very desirable feature which is not true for the block bootstrap; see (2.5). These adaptivity results are not only asymptotically relevant but can be well exploited in finite sample simulations; see Section 4.2.

The AR-sieve bootstrap is not only very accurate for variance estimation, Choi and Hall (2000) show a 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

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

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

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

Define

$$\begin{aligned} &\hat{a}(1-q) \\ &\quad = \mathbb{P}^{*\text{AR-S}}[\hat{\theta}^{*\text{AR-S}} - \hat{r}_{1-q/2}^{*\text{AR-S}} \\ &\qquad \le \theta^{*\text{AR-S}} \le \hat{\theta}^{*\text{AR-S}} - \hat{r}_{q/2}^{*\text{AR-S}}], \end{aligned} \tag{3.5}$$

measuring actual coverage on nominal level $1-q$ for the second level bootstrap based on the first level bootstrapped data ($\theta^{*\text{AR-S}}$ is a constant depending only on $X_1, \ldots, X_n$). Then, consider

$$\hat{s}_{1-\alpha} = \hat{a}^{-1}(1-\alpha)$$

[the $(1-\alpha)$-quantile of $\hat{a}$ viewed as a cdf], which corrects the nominal coverage level $1-\alpha$ to $\hat{s}_{1-\alpha}$. (In Section 5.3, Figure 10 illustrates the correction of the nominal coverage level $1-\alpha$ to $\hat{s}_{1-\alpha}$.) Now use

$$[\hat{\theta} - \hat{r}_{\{1-(1-\hat{s}_{1-\alpha})/2\}}, \hat{\theta} - \hat{r}_{\{(1-\hat{s}_{1-\alpha})/2\}}] \tag{3.6}$$

as a two-sided, double bootstrap confidence interval for θ with nominal coverage level $1-\alpha$. As shown by Choi and Hall (2000), this interval is second-order correct. Note that explicit (difficult) variance estimation with dependent data for Studentizing is not necessary. Choi and Hall (2000) report from a simulation study that this second-order interval can bring very substantial improvements and has "never" been found significantly worse than the first-order construction.

3.3 Choosing the Approximating Autoregressive Order

We propose the AR-sieve approximation in conjunction with the minimum AIC model selection procedure with Gaussian innovations. Shibata (1980) has shown optimality of the AIC for prediction in AR(∞) models. Moreover, (3.4) and its preceding formula, which are both based on AIC, explain why the criterion is a good choice for variance estimation of $\hat{\theta} = \overline{X}_n$.

Analogous to the problem of choosing an optimal blocklength for the block bootstrap, the optimal autoregressive order generally depends on the true underlying process, the statistic to be bootstrapped and the purpose for what the bootstrap is used. The AIC has the nice property of automatically selecting higher orders for more strongly dependent models. Nothing is known of 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 for AR-residuals. This is in contrast to bandwidth-type tuning parameters like the blocklength in Section 2.3, which have no good interpretation and are not easy to "back-test" on the data. Our empirical experience is that the choice of an approximating autoregressive order is quite *insensitive* with respect to the performance of the AR-sieve bootstrap, provided the chosen order is reasonable.

4. BLOCK AND AR-SIEVE BOOTSTRAP IN ACTION

4.1 Total Ozone Series from Arosa

We consider here the world's longest series of monthly total ozone measurements, from Arosa, Switzerland, during the period 1926–1997. It is an important source for assessing the ozone depletion in the midlatitudes of the northern 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 $\{O_t\}$ exhibit big seasonal effects which can be explained very well. Assuming fixed monthly effects β_i $(i = 1, \dots, 12)$ with $\sum_{i=1}^{12} \beta_i = 0$, we deseasonalize the series by preliminary smoothing with a running mean $X_t = \sum_{i=-6}^{6} c_i \cdot O_{t-i}$ with $c_i = 1/12$ $(i = -5, \dots, 5)$ and $c_i = 1/24$ $(i = -6, 6)$. Figure 2 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 2. An additional question is about the ozone variability around a varying trend whose estimate is also displayed in Figure 2. Here we use time series bootstraps to assess statistical variability of these trend and variability smoothers and to answer the questions whether trend and/or variability change significantly over time.

Consider the basis model

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

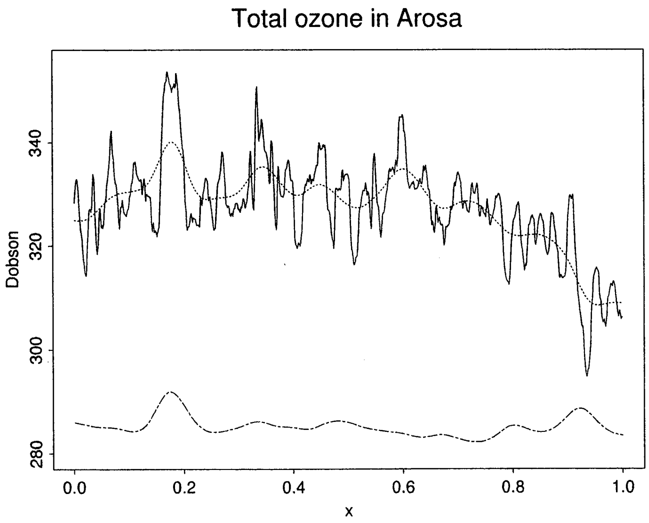

FIG. 2. *Total deseasonalized ozone measurements (solid line), mean trend smoother $\hat{m}(\cdot)$ (dotted line) and magnitude of smoother $\hat{s}(\cdot)$ for changing variability (dashed line). Time, ranging from January* 1927 *to June* 1997, *is rescaled to* (0, 1].

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 $\operatorname{Var}(Z_t) = 1$. What Figure 2 shows are kernel estimates of $m(\cdot)$ and $s(\cdot)$, defined as follows. For the mean function,

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

where K is the standard Gaussian kernel and the bandwidth $h = 0.024$ is chosen by eye.

For the scale function, build the transformed values

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

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 $\gamma_t = \mathbb{E}[\log(Z_t^2)] + \log(s^2(t/n))$. Transforming back by exponentiating and estimating $\exp(\mathbb{E}[\log(Z_t^2)])$ by $(n^{-1}\sum_{t=1}^{n}[(X_t - \hat{m}(t/n))^2/\exp(\hat{\gamma}_t)])^{-1}$ (using that $\mathbb{E}|Z_t|^2 = 1$) 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. We apply some bootstraps to the residual process $\hat{Z}_t = (X_t - \hat{m}(t/n))/\hat{s}(t/n)$ yielding $Z_t^*, t = 1, \dots, n$. Bootstrapping from the null-distribution is then done as

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

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

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

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

$$\text{(4.1)} \quad \hat{m}^{*H_1}(\cdot) \text{ based on } X_1^{*H_1}, \dots, X_n^{*H_1} \text{ for } H_1,$$

$$\text{(4.2)} \quad \hat{s}^{*H_2}(\cdot) \text{ based on } X_1^{*H_2}, \dots, X_n^{*H_2} \text{ for } H_2.$$

The construction of the resampled noise process Z_t^* $(t = 1, \dots, n)$, 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 $\ell = 9 \approx n^{1/3}$ (according to a simple rule, see Section 2.3) 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 2.3). Figures 3 and 4 show the estimates $\hat{m}(\cdot)$ and $\hat{s}(\cdot)$ together with 19 bootstrap replicates each from the estimates in (4.1) and (4.2), respectively. They display

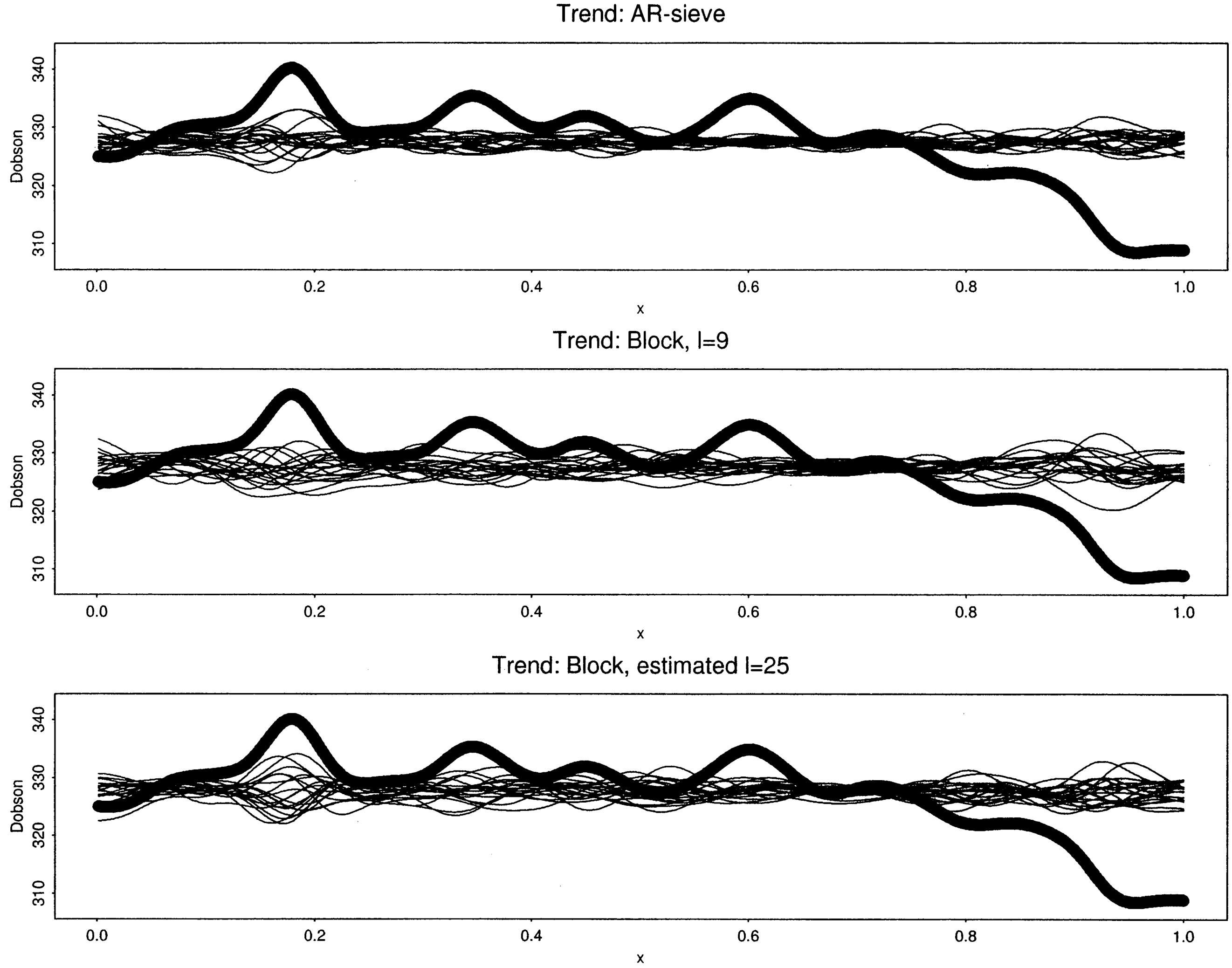

FIG. 3. *Mean trend estimates*: 19 *bootstrapped estimators* $\hat{m}^{*H_1}(\cdot)$ *under the hypothesis* H_1 *with constant trend* (*fine lines*), *the estimator* $\hat{m}(\cdot)$ *based on original data* (*bold line*); *AR-sieve bootstrap with AIC estimated order* 29, *block bootstrap with* $\ell = 9$ *and estimated* $\hat{\ell} = 25$.

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 20 curves, corresponding to a 5% significance level for testing. Of course, a more formal construction of acceptance regions for say two-sided testing of H_1 and H_2 would be possible.

All three bootstrap methods lead to similar conclusions, increasing confidence about the appropriateness of the resampling methods displayed in Figures 3 and 4. It is very valuable to have both the AR-sieve and block bootstrap as tools in a practical example. Regarding the mean trend, there is clear evidence for a decreasing behavior as time progresses. 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.

4.2 AR-Sieve versus Block Bootstrap for Simulated Series

For comparing the two bootstraps, we also consider simulation experiments with two different processes but with the statistic being in both cases the sample median $\hat{\theta} = \text{med}(X_1, \dots, X_n)$, representing a simple nonlinear estimator. The sample sizes are $n = 512$. Furthermore, the tuning parameters are chosen by the minimal AIC for the AR-sieve; $\ell = 8 = n^{1/3}$ according to a simple rule having the optimal asymptotic rate for variance estimation, and $\hat{\ell}$ from Bühlmann and Künsch (1999) for the block bootstrap variance as indicated in Section 2.3.

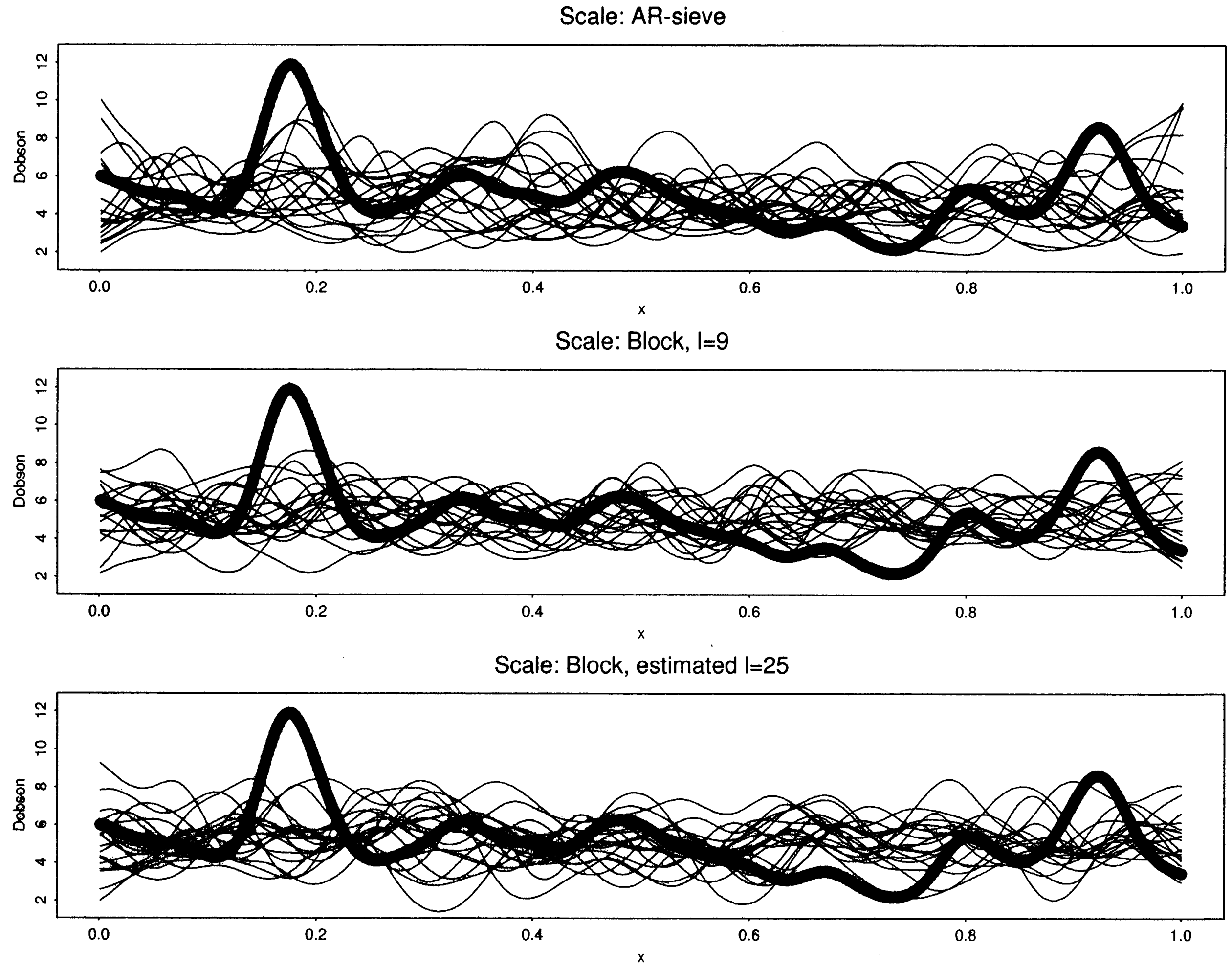

FIG. 4. *Estimates for scale*: 19 *bootstrapped estimators* $\hat{s}^{*H_2}(\cdot)$ *under the hypothesis* H_2 *with constant scale* (*fine lines*), *the estimator* $\hat{s}(\cdot)$ *based on original data* (*bold line*); *AR-sieve bootstrap with AIC estimated order* 29, *block bootstrap with* $\ell = 9$ *and estimated* $\hat{\ell} = 25$.

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

$$X_t = -0.8X_{t-1} - 0.5\varepsilon_{t-1} + \varepsilon_t, \tag{4.3}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d sequence, independent of $\{X_s;\ s < t\}$, $\varepsilon_t \sim t_6$. This model is representable as an AR(∞)-process as in (3.1).

Figure 5 displays the quality of bootstrap approximations for the sample median in model (4.3). The AR-sieve bootstrap outperforms the block bootstrap very clearly. Estimation of the blocklength improves a bit upon the fixed blocklength $\ell = 8 = n^{1/3}$: typical values of $\hat{\ell}$ in the 100 simulations are 22, 16 and 30 corresponding to the median value, lower quartile and upper quartile, respectively. The better performance of the AR-sieve bootstrap is not so surprising: we exploit here the advantages discussed in Section 3.2. The result here indicates quantitatively the gain in a case where the true underlying process is not a finite-order AR-model, and hence not an element of the approximating sieve (for any finite sample size), but is representable as AR(∞) as in (3.1). 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 pseudoperiodic decay which is true for the model in (4.3). 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,

$$\begin{aligned} X_t &= (0.5 + 0.9\exp(-X_{t-1}^2))X_{t-1} \\ &\quad - (0.8 - 1.8\exp(-X_{t-1}^2))X_{t-2} + \sigma_t\varepsilon_t, \\ \sigma_t^2 &= 0.5 + 0.1X_{t-1}^2 + 0.05\sigma_{t-1}^2\mathbb{1}_{[X_{t-1}\le 0]} \\ &\quad + 0.5\exp(-\sigma_{t-1}^2)\mathbb{1}_{[X_{t-1}>0]}, \end{aligned} \tag{4.4}$$

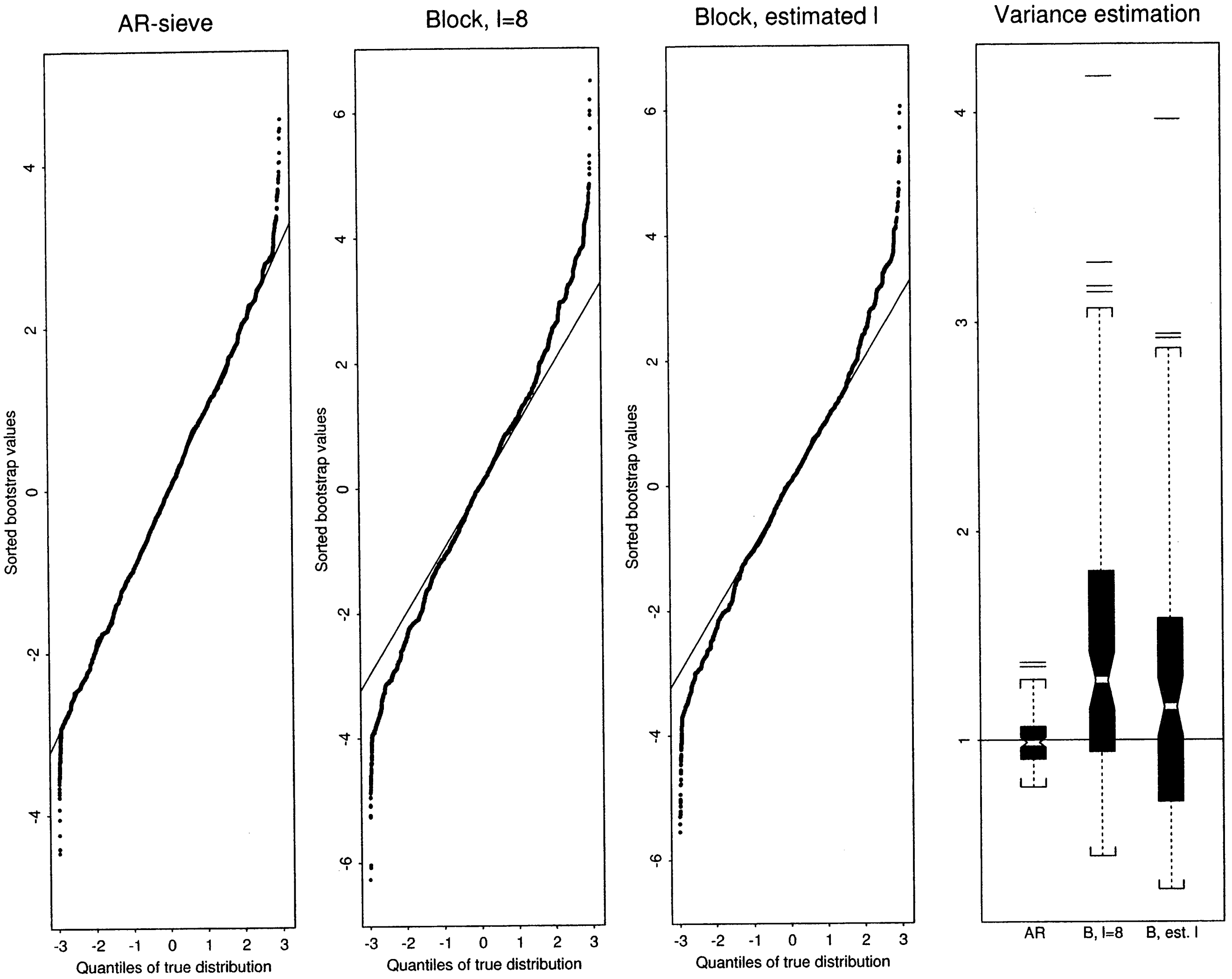

FIG. 5. *Linear model* (4.3), $n = 512$; *bootstrap distribution and variance estimation of* $(\hat{\theta} - \mathbb{E}[\hat{\theta}])/\sigma_n$ *by* $(\hat{\theta}^* - \mathbb{E}^*[\hat{\theta}^*])/\sigma_n$ *for* $\hat{\theta} = \mathrm{med}(X_1, \dots, X_n)$ $[\sigma_n = (\mathrm{Var}(\hat{\theta}))^{1/2}]$; (*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.*

where $(\varepsilon_t)_{t \in \mathbb{Z}}$ is an i.i.d sequence, independent of $\{X_s;\ s < t\}$, $\varepsilon_t \sim t_6/\sqrt{1.5}$. This process is not representable as an AR(∞) as in (3.1).

Figure 6 displays the quality of bootstrap approximations for the sample median in model (4.4). The AR-sieve bootstrap, which is not asymptotically consistent due to the nonlinearity of the model in (4.4), exhibits a clear bias, and the block bootstrap is superior. As in the linear case (4.3), using the estimated blocklength $\hat{\ell}$ improves upon the fixed blocklength $\ell = 8 = n^{1/3}$: typical values of $\hat{\ell}$ in the 100 simulations are 11, 7 and 14, corresponding to the median value, lower quartile and upper quartile, respectively. The results in Figure 6 again give quantitative insights about the gain when using the block bootstrap in this nonlinear model.

4.3 Comparison with Subsampling

Subsampling blocks is a very general technique for estimating moments or the distribution of an estimator $\hat{\theta}$. The basic idea is to compute an estimator $\hat{\theta} = h_n(X_1, \dots, X_n)$ over many subsamples of ℓ consecutive observations (blocks)

$$\hat{\theta}_{\ell,t} = h_\ell(X_{t-\ell+1}, \dots, X_t), \quad t = \ell, \dots, n.$$

Distribution and variance approximations with subsampling are then constructed as

$$(n - \ell + 1)^{-1} \sum_{t=\ell}^{n} \mathbb{1}_{[a_\ell(\hat{\theta}_{\ell,t} - \hat{\theta}) \le x]} \approx \mathbb{P}[a_n(\hat{\theta} - \theta) \le x],$$

$$(n - \ell + 1)^{-1} \sum_{t=\ell}^{n} (a_\ell(\hat{\theta}_{\ell,t} - \hat{\theta}))^2 \approx a_n^2 \mathrm{Var}(\hat{\theta})$$

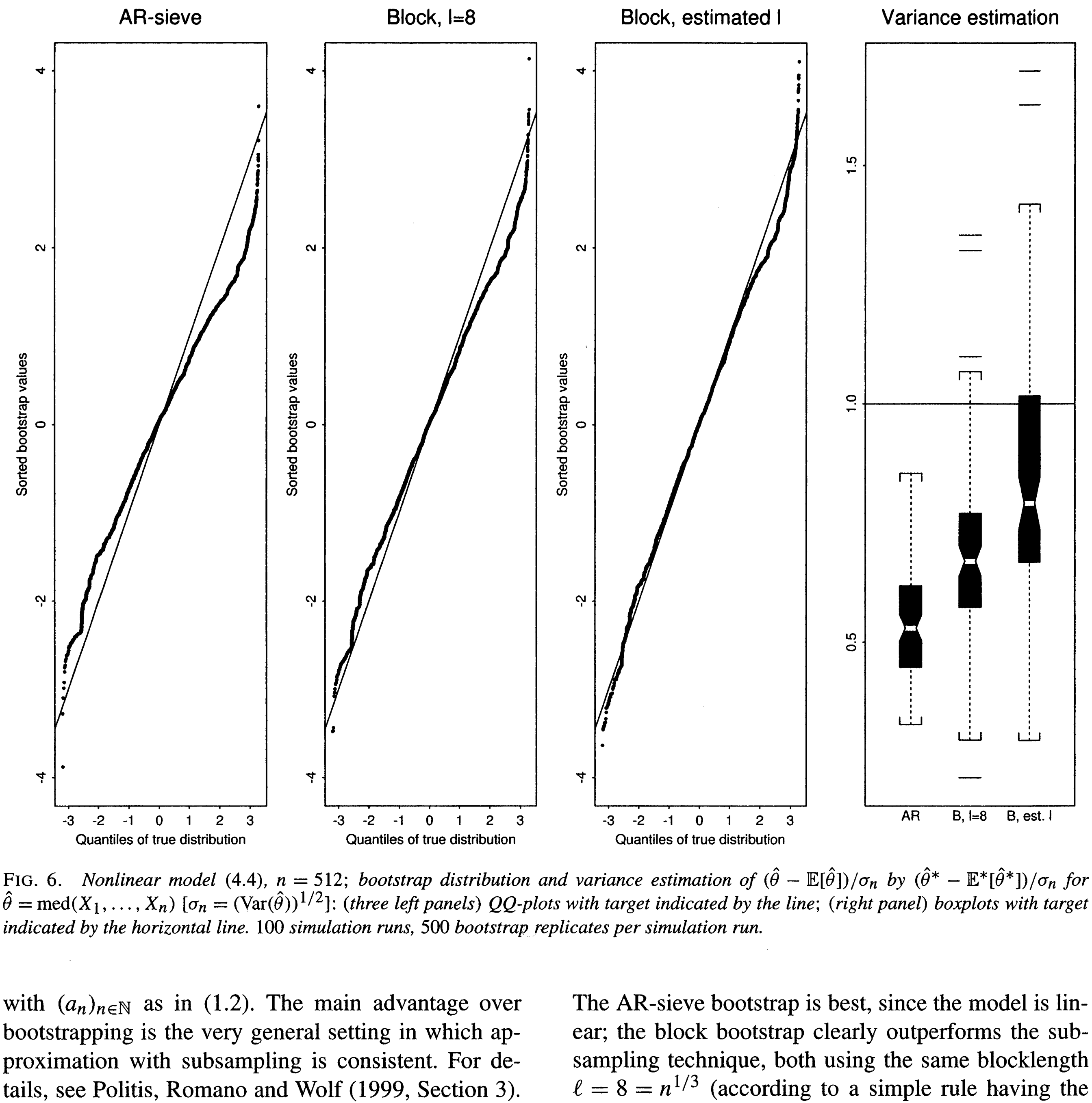

FIG. 6. *Nonlinear model* (4.4), $n = 512$; *bootstrap distribution and variance estimation of* $(\hat{\theta} - \mathbb{E}[\hat{\theta}])/\sigma_n$ *by* $(\hat{\theta}^* - \mathbb{E}^*[\hat{\theta}^*])/\sigma_n$ *for* $\hat{\theta} = \mathrm{med}(X_1, \dots, X_n)$ [$\sigma_n = (\mathrm{Var}(\hat{\theta}))^{1/2}$]: (*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.*

with $(a_n)_{n \in \mathbb{N}}$ as in (1.2). The main advantage over bootstrapping is the very general setting in which approximation with subsampling is consistent. For details, see Politis, Romano and Wolf (1999, Section 3). However, computing an estimator $\hat{\theta}$ on subsamples of much smaller size $\ell \ll n$ and scaling up to its behavior of the original sample size n can be problematic when $\hat{\theta}$ is highly nonlinear and sample n is not extremely large. A simple but impressive example is the empirical lag(1) autocorrelation

$$\hat{\theta} = \hat{\rho}(1) = \hat{R}(1)/\hat{R}(0)$$

as in Example A from Section 2.1.

Figure 7 displays the results for variance estimation of $\hat{\rho}(1)$ in model (4.3) with sample size $n = 512$. The AR-sieve bootstrap is best, since the model is linear; the block bootstrap clearly outperforms the subsampling technique, both using the same blocklength $\ell = 8 = n^{1/3}$ (according to a simple rule having the correct asymptotic rate for both methods). Generally, it is not advisable to use subsampling when the bootstrap is known to be consistent. Subsampling is an interesting tool for complicated procedures $\hat{\theta}$ (on a large data set) where bootstrap methods potentially fail.

5. VARIABLE LENGTH MARKOV CHAIN SIEVE BOOTSTRAP FOR STATIONARY CATEGORICAL TIME SERIES

Sieve approximation is also successful for general stationary processes $(X_t)_{t \in \mathbb{Z}}$ with values in a categori-

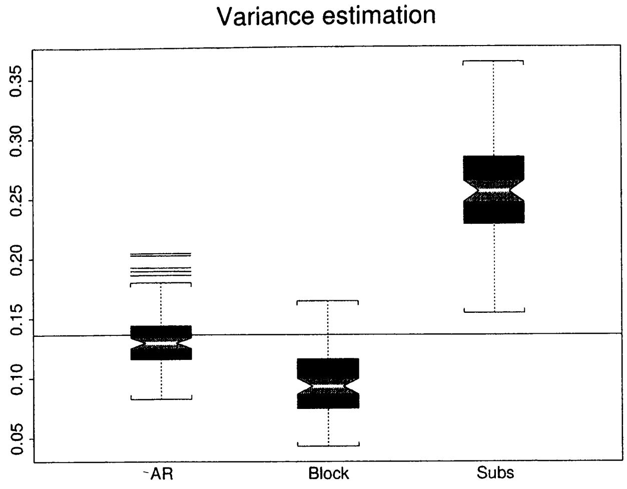

FIG. 7. *Linear model* (4.3), $n = 512$: *variance estimation of* $n\mathrm{Var}(\hat{\theta})$ *for the estimated lag*-1 *autocorrelation* $\hat{\theta} = \hat{\rho}(1)$ (*horizontal line*); *"AR", "Block" and "Subs" denote AR-sieve, block bootstrap and subsampling, respectively*; *the latter two with blocklengths* $\ell = 8$. 100 *simulation runs and* 500 *replicates per simulation run for the bootstraps.*

cal, finite space $\mathcal{X}$. For example, data from a DNA sequence with values in the set $\{A, C, T, G\}$ build a categorical time series. We consider the sieve of so-called variable length Markov chains (VLMC) for approximating $\mathcal{X}$-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 potentially high order whose time-homogeneous transition probabilities depend on a *variable* number ℓ of lagged values,

$$\begin{aligned}&\mathbb{P}[X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots]\\ &\quad = \mathbb{P}[X_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_{t-\ell} = x_{t-\ell}]\end{aligned}$$

for all $x_t \in \mathcal{X}$, 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 can always embed a VLMC in a full Markov chain of order p, but with the additional *structure of a variable length memory*. It implies that some transition probabilities of the embedding Markov chain are the same; that is, some rows in the matrix of Markov transition probabilities are identical. The variable length memory is essential for parsimony while still being flexible: it is an attractive approach to dealing intelligently with the curse of dimensionality which is heavily present in full Markov chains. The main difficulty is the estimation of $\ell(\cdot)$, the structure of the variable length memory, from data. Since $\ell(\cdot)$ is discrete, the task can be viewed as a highly complex model selection problem among an enormous number of possible candidate models.

A version of the tree structured context algorithm (Rissanen, 1983) can be used for estimating $\ell(\cdot)$ and the set of transition probabilities. The exact description of the algorithm as used here can be found in Bühlmann and Wyner (1999) or Bühlmann (2002). It yields a consistent estimate $\hat{P}_{n;\mathrm{VLMC}}$ for the distribution of suitably regular processes which is not necessarily a VLMC.

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

$$X_1^{*\mathrm{VLMC\text{-}S}}, \ldots, X_n^{*\mathrm{VLMC\text{-}S}} \sim \hat{P}_{n;\mathrm{VLMC}}. \tag{5.1}$$

We briefly describe in Section 5.2 how this can be computed using the software *R*. Having the bootstrap sample in (5.1), we proceed by using the plug-in principle, exactly as in (3.3).

5.1 Range of Applicability and Accuracy

The VLMC-sieve bootstrap is designed to be consistent for data-generating stationary categorical processes which are short-range dependent (e.g., summable mixing coefficients). Asymptotically, the context algorithm for fitting VLMC's automatically selects larger models (or finds the true VLMC model) as $n \to \infty$. Consistency as in (1.2) then holds for general estimators of the form (2.3) defined in Section 2.1. More details are given in Bühlmann (2002).

For variance estimation, the VLMC-sieve bootstrap has good convergence rates: if the mixing coefficients decay exponentially fast as separation lags increase, and if the data-generating process is suitably regular (not necessarily a VLMC),

$$\begin{aligned}&n\mathrm{Var}^{*\mathrm{VLMC\text{-}S}}(\hat{\theta}^{*\mathrm{VLMC\text{-}S}}) - n\mathrm{Var}(\hat{\theta})\\ &\quad = O_P(n^{-1/2+\varepsilon}) \quad \text{for any } \varepsilon > 0,\end{aligned} \tag{5.2}$$

where $\hat{\theta} = (n - m + 1)^{-1}\sum_{t=m}^{n} f(X_{t-m+1}, \ldots, X_t)$ with $f\colon \mathcal{X}^m \to \mathbb{R}$ ($m \in \mathbb{N}$) (cf. Bühlmann, 2002). The bound in (5.2), achieved in a *data-driven* way, is much better than (2.5) for the block bootstrap.

Double VLMC bootstrapping and construction of a calibrated confidence interval can be done analogously to (3.6), aiming for higher order coverage properties.

5.2 Computation and Tuning Parameter Selection

The context algorithm and the VLMC-sieve bootstrap are implemented in the statistical computing language *R*, freely available from the download section of http://www.rproject.org/. The exact commands in *R* look as follows:

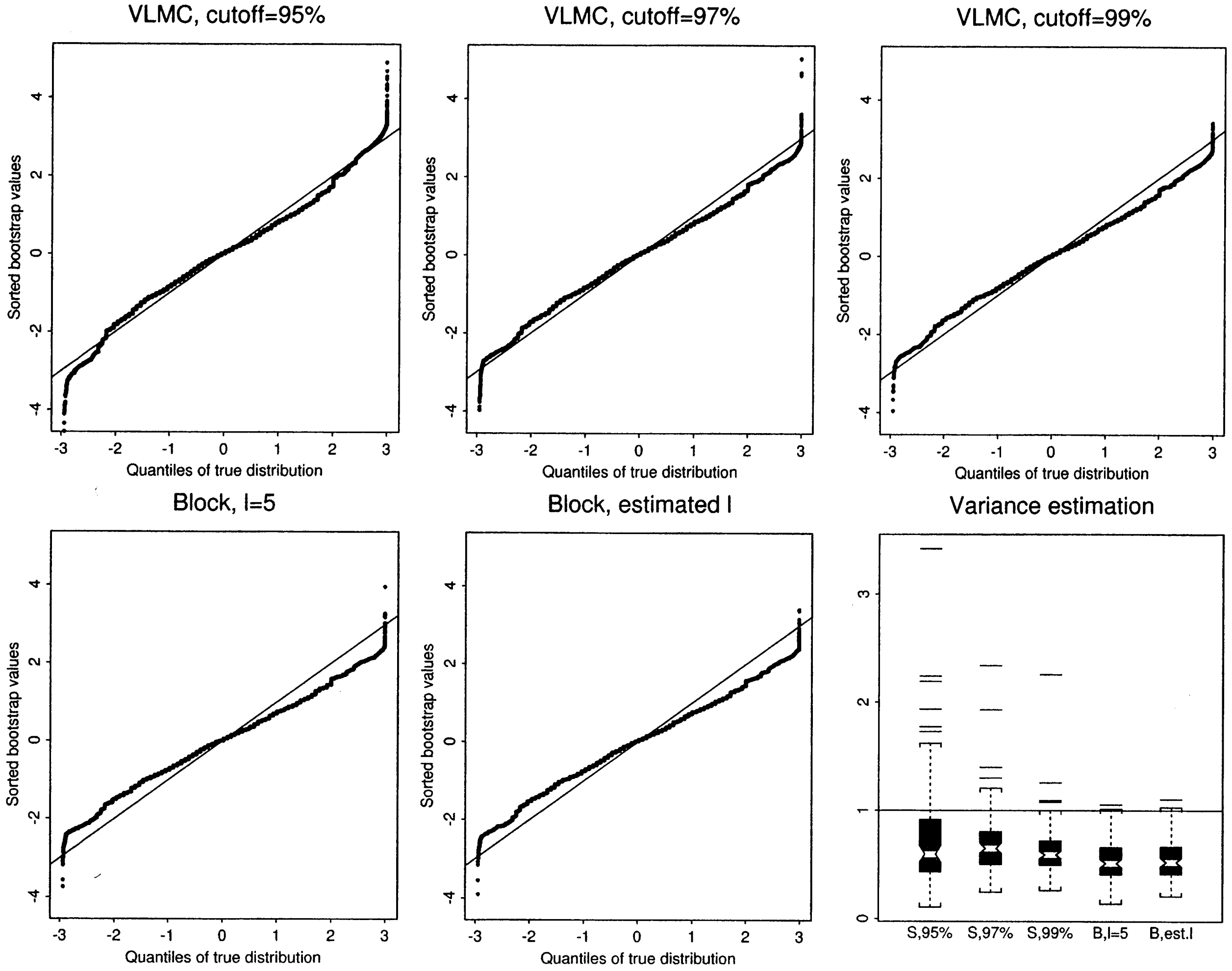

FIG. 8. *Bootstrap distribution and variance estimation of* $(\overline{X}_n - \mathbb{E}[X_t])/\sigma_n$ *by* $(\overline{X}_n^* - \mathbb{E}^*[\overline{X}_n^*])/\sigma_n$ *for* $n = 128$ [$\sigma_n = (\mathrm{Var}(\overline{X}_n))^{1/2}$]: *the target is indicated by the line. VLMC-sieve bootstrap with cutoff values* $\chi^2_{1;\alpha}/2$ *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.*

```
> library(VLMC)
> fit <- vlmc(series,cutoff=C)
```

This is the VLMC fit of the data with the context algorithm. Thereby, the so-called cutoff value is used as a tuning parameter. The default **cutoff** is $C = \chi^2_{\mathrm{card}(\mathcal{X})-1;0.95}/2$, half of the 95% quantile of a χ^2-distribution with $\mathrm{card}(\mathcal{X}) - 1$ degrees of freedom.

```
> simulate.vlmc(fit, resample size)
```

This is the VLMC-sieve bootstrap sample.

The cutoff tuning parameter mentioned above characterizes a computationally efficient selection of a VLMC model structure. A simple data-driven version for selecting the **cutoff** can be implemented by minimizing the AIC statistic: the AIC statistics of the fitted model **fit** is given by

```
> AIC(fit)
```

More on-line help is implemented in *R* with library (VLMC) and additional details can be found in Bühlmann (2002).

5.3 VLMC-Sieve versus Block Bootstrap for Simulated Series

The differences in variance estimation between (2.5) and (5.2) can be well exploited in finite-sample problems, and the gain of the VLMC-sieve bootstrap is often, although not always, substantial.

We consider a simulated example where

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

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d. innovation sequence, independent of $\{Y_s;\ s < t\}$, $\varepsilon_t \sim t_6$. Of interest here 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 nonsparse 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) $\hat{\theta}$ is the probability of the five-tuple $(1, 1, 1, 1, 1)$ in an estimated VLMC model.

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 constructed with the plug-in rule in (3.3). Block bootstrapping for (S1) requires no vectorization step as described in Section 2.1. The estimator (S2) is an example where redesigning the computation for block bootstrapping with vectorization is almost impossible: the VLMC-estimator (S2) is computationally implemented with input being a sequence of categorical variables. The only feasible way for block bootstrapping (S2) is simply to ignore the vectorization step. 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 simple choice of 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 2.3. Figure 8 displays the quality of distribution estimation for the estimator (S1). For distribution estimation, the VLMC-sieve bootstrap is better than the block bootstrap for a whole range of cutoff tuning parameters, at least not too far out in the tails [not that the true distribution is close to $\mathcal{N}(0, 1)$ and hence quantiles around ±2 are often of interest]. For variance estimation, the advantage of the VLMC method is less pronounced, but still present: the VLMC-sieve bootstrap with the best tuning parameter has 20.5% lower mean squared error for variance estimation than the best block bootstrap. Using the estimated blocklength $\hat{\ell}$ improves a bit upon the fixed blocklength $\ell = 5 \approx n^{1/3}$; typical values of $\hat{\ell}$ in the 100 simulations are 8, 6 and 9, corresponding to the median value, lower quartile and upper quartile, respectively.

Figure 9 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 in this case, the VLMC-sieve is better than the block bootstrap. The VLMC method produces a few outliers for variance estimation which indicates a small chance that the VLMC-sieve approximation can be bad for the complicated estimator (S2).

Figures 8 and 9 are representative of other situations with exponentially decaying dependence structure: very often, the VLMC-sieve bootstrap is better than the block bootstrap, the latter being also more sensitive to the specification of the blocklength parameter. In practice, a procedure which is insensitive to the choice of tuning parameters is highly desirable. An example where the block bootstrap is better than the VLMC method is given in Bühlmann (2002): there the underlying model exhibits only dependencies over neighboring values [lag(1)-dependence], which is generally favorable for the block bootstrap.

We also examine construction of a 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 consider first-order accurate block and VLMC-sieve bootstraps and corrections thereof with a version of BC$_a$ for the block bootstrap (cf. Götze and Künsch, 1996) and the double bootstrap for the VLMC method, analogously to the AR-sieve scheme in section 3.2, formula (3.6) (with the same tuning parameter for the first- and second-level bootstrap). The tuning parameters of the methods correspond to the upper left and lower right panels in Figure 8. Due to the discreteness of the observations X_t (latticeness of the problem), correction of confidence intervals does not seem worthwhile from an asymptotic point of view. However, there still may be some considerable gain to employ corrections for finite samples. For a related discussion see Hall (1987) and Woodroofe and Jhun (1988). Coverage probabilities of confidence intervals with median and mean absolute deviation of their lengths are given in Table 1. The non-Markovian

TABLE 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*

	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

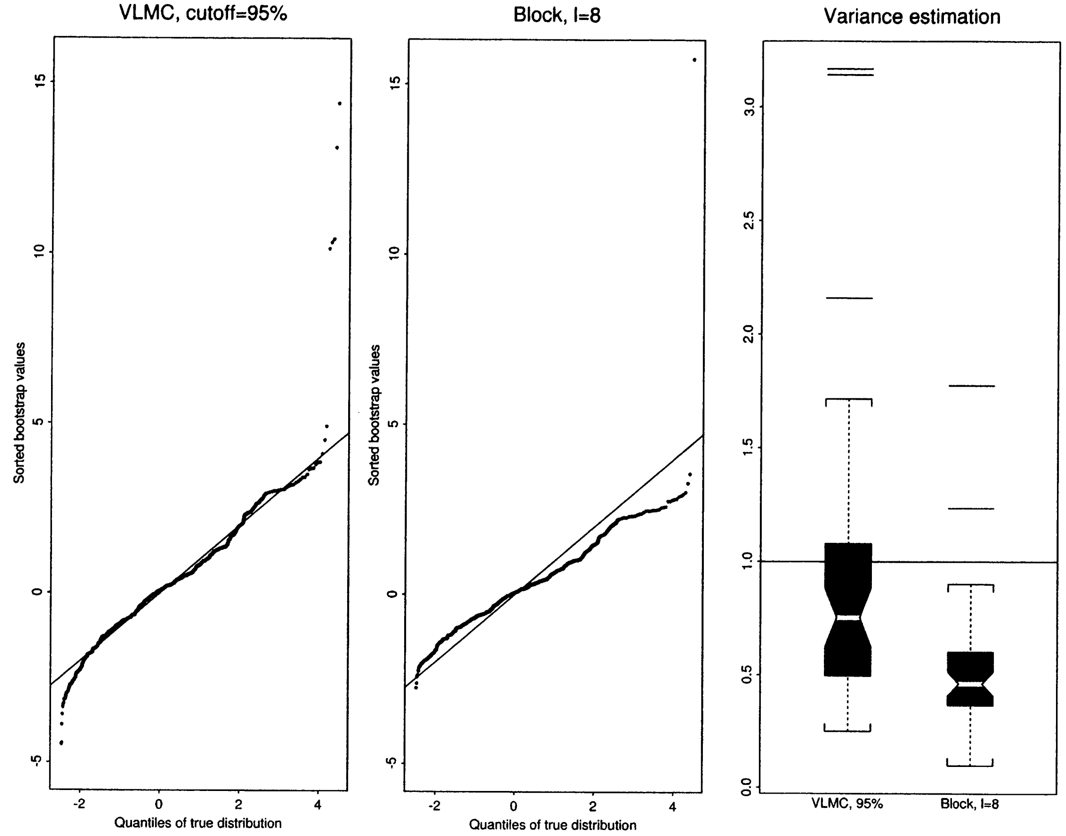

FIG. 9. *Bootstrap distribution and variance estimation of* $(\hat{\theta} - \mathbb{E}[\hat{\theta}])/\sigma_n$ *by* $(\hat{\theta}^* - \mathbb{E}^*[\hat{\theta}^*])/\sigma_n$ *for* $\hat{\theta}$ *from (S2) and* $n = 512$ [$\sigma_n = (\mathrm{Var}(\hat{\theta}))^{1/2}$]: (*two left panels*) *QQ-plots 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.*

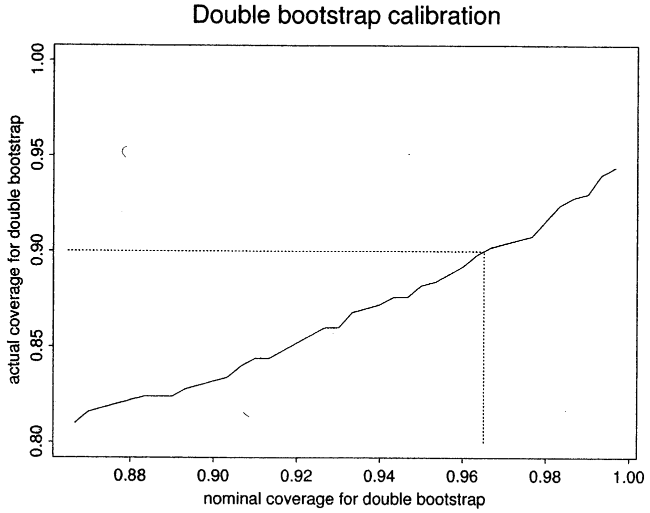

FIG. 10. *Double VLMC bootstrap calibration for one typical sample of size* $n = 128$: (*x-axis*) *nominal coverage level for second-level bootstrap based on first-level bootstrapped data*; (*y-axis*) *corresponding actual coverage level. These are the quantities* $1 - q$ *and* $\hat{a}(1 - q)$ *in* (3.5); *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-level and* 500 *second-level bootstrap replicates.*

model used in the simulation and the small sample size $n = 128$ make interval estimation difficult and explain the generally poor coverage results. For the block bootstrap, the BC_a method increases performance with weak significance compared to the first-order block interval. In comparison with the first-order VLMC and with any of the block methods, the double VLMC bootstrap improves with strong significance upon coverage. On average over the 100 simulations, it corrects the nominal 90% to the 97.3% coverage level; a calibration for one typical sample is shown in Figure 10, yielding the corrected coverage level 96.7%.

6. A LOCAL BOOTSTRAP FOR CONDITIONAL MEAN ESTIMATES

The sieve and block bootstraps give reasonable results for a variety of estimators $\hat{\theta}$, whenever the data-generating process belongs to an appropriate range for the various bootstraps, as discussed above. A bit surprisingly, some bootstraps based on independent resampling can be used for the class of nonparametric estimators $\hat{\theta}$ having slower rate of convergence than $1/\sqrt{n}$: for example, $\hat{\theta}(\cdot)$ a (kernel) smoother of the conditional expectation $\theta(\cdot) = \mathbb{E}[X_t \mid X_{t-1} = \cdot]$ of a stationary process.

We focus on interval estimation with a so-called local bootstrap for the conditional expectation $\theta(x) = \mathbb{E}[X_t \mid X_{t-1} = x]$, $x \in \mathbb{R}$, 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 $\mathbb{E}[f(X_t) \mid X_{t-i_1} = x_1, \ldots, X_{t-i_p} = x_p]$ for a specified set of p lagged indices $t - i_1, \ldots, t - i_p$ and $f\colon \mathbb{R} \to \mathbb{R}$ (in practice, p most often smaller or equal to 2). Given data $X_1, \ldots, X_n$, consider the kernel estimate

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

with bandwidth h.

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

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

where $\hat{F}_{x,b}(\cdot) = \sum_{t=2}^n W_{t,b}(x) \mathbb{1}_{[X_t \le \cdot]} / \sum_{t=2}^n W_{t,b}(x)$ is an estimate of the conditional cumulative distribution of X_t given $X_{t-1} = x$; b is a (pilot) bandwidth and $W_{\cdot,\cdot}(\cdot)$ as above. Thus, the resampling is constructed independently with the estimated $\{\hat{F}_{x,b}(\cdot); x \in \mathbb{R}\}$ 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 considered by Neumann and Kreiss (1998) and Paparoditis and Politis (2000).

The local bootstrap works because the asymptotic distribution of the kernel estimator $\hat{\theta}_h(x)$ is Gaussian, depending 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; see Robinson (1983). The local bootstrap is able to estimate consistently all these unknowns. This is only asymptotically true and for any finite sample size n, already 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 cannot pick up dependencies beyond the conditional distribution of X_t given X_{t-1}. This disadvantage does not occur with bootstraps designed for dependent data, for example, the block bootstrap: Accola (1998) proves a better rate of estimating $\mathrm{Var}(\hat{\theta}_h(x))$ with the block than the local bootstrap.

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

6.1 Range of Applicability and Selection of the Tuning Parameter

The local bootstrap is proven to be consistent whenever $(X_t)_{t\in\mathbb{Z}}$ is a short-range dependent process (cf. Paparoditis and Politis, 2000, and Ango Nze, Bühlmann and Doukhan, 2002).

The tuning parameter of the local bootstrap is the pilot bandwidth b. A simple approach is to choose $b = h$, where h is the prechosen bandwidth of the estimator $\hat{\theta}_h(x)$. When b is of larger order than h, an asymptotically nonnegligible bias $\mathbb{E}[\hat{\theta}_h(x)] - \theta(x)$ can be estimated with the local bootstrap (cf. Paparoditis and Politis, 2000). The pilot bandwidth plays a role in estimating the conditional distribution of X_t given X_{t-1}: this task is relatively easy for two-dimensional distributions. The procedure seems not very sensitive to specification of this pilot bandwidth.

6.2 Local versus Block Bootstrap for Simulated Series

From a finite sample 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 simulation experiment which deals with a bilinear model,

$$X_t = 0.5\varepsilon_{t-1} X_{t-1} + \varepsilon_t, \tag{6.1}$$

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

(M1) ε_t i.i.d. $\sim$ Uniform($\{$-1,1$\}$), that is, $\mathbb{P}[\varepsilon_t = 1] = \mathbb{P}[\varepsilon_t = -1] = 1/2$;

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

Both models (M1) and (M2) exhibit weak forms of dependence. We found empirically that estimation of Var$(\hat{\theta}_h(x))$ is harder in the discrete innovation model (M1) than in (M2).

Figure 11 displays results for estimating Var$(\hat{\theta}_h(x))$, with standard Gaussian kernel K and reasonable bandwidth $h = 0.25$. Sample size is $n = 512$. Graphical detection of relevant differences is difficult. A more quantitative description 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, 0.5"): the two-sided paired Wilcoxon test favors "B, 5" with a p-value of 0.002 for the null hypothesis of equal MSE. Comparing any of the local with any of the other block bootstraps with $\ell = 1, 3, 8$ yields 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 the difference is nonsignificant. Excluding "B, 8" with unreasonably large blocklength (having a significant disadvantage with respect to local bootstraps), any of the local compared with any of the other block bootstraps with $\ell = 1, 3, 5$ yields no significant difference. For this specific bilinear model with weak degree of dependence we conclude the following. In the easier case (M2), the local and block bootstraps are equally good (when excluding the unreasonable blocklength $\ell = 8$). This contrasts a bit with the harder case (M1) where the block bootstrap is always as good as local bootstraps and even better, provided we have a good rule for choosing a blocklength around $\ell = 5$. We expect the block bootstrap to be clearly better than the local bootstrap whenever the data exhibits stronger degree of dependence.

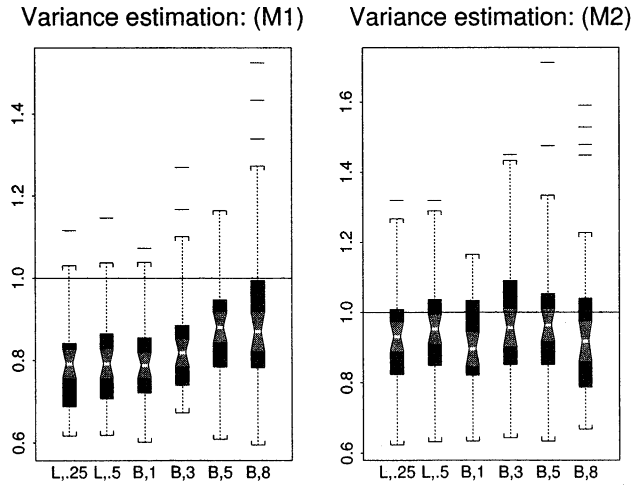

FIG. 11. *Bootstrap variance estimates* Var*$(\hat{\theta}_h^*(x))$/Var$(\hat{\theta}_h(x))$ *at* 80% *quantile of marginal distribution, namely* $x = 1.60$ *and* $x = 0.76$ *for* (*M*1) *and* (*M*2), *respectively* (*the target is indicated by the horizontal line*): "L" *for local bootstrap with pilot bandwidths* 0.25 *and* 0.5; "B" *for block bootstrap with blocklengths* 1, 3, 5 *and* 8. *Sample size is* $n = 512$.

7. CONCLUSIONS

Among the block bootstrap, two types of sieve bootstrap and a local resampling scheme, the block bootstrap is the most general method. A further advantage is its 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 sample mimicking the data-generating process: it is not 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. A prevectorization of the data is highly recommended, but the bootstrapped estimator and its computing routine may then need to be redesigned. As a general nonparametric scheme, the block bootstrap may be outperformed in various niches of stationary time series, for example, for linear time series (see Section 3) and for categorical processes (see Section 5). Second-order accuracy for a confidence interval has been justified with the approach of Studentizing and BC$_a$ correction (in the case of noncategorical time series); the latter was found to yield marginal improvement in a simulated example (we did not consider the former). Double bootstrapping does not seem promising since the block bootstrap in the first iteration corrupts dependence where blocks join.

Sieve bootstraps in general resample 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, like the AR- or VLMC-sieve schemes, are expected to adapt to the degree of dependence: their accuracy improves as the degree of dependence decreases; see (3.4) and (5.2). This is not the case with the block bootstrap, as seen from (2.5). 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 (3.1). The method is easy to implement, due to the simplicity of fitting an AR model.

The VLMC-sieve bootstrap is often best for categorical processes. The disadvantage is the difficulty of doing the resampling: the context algorithm which is used for this task is publicly available in R, which should help to overcome most of the implementational burdens. The algorithm is computationally fast using only $O(n \log(n))$ essential operations. Double bootstrapping was successful in a simulated example.

The local bootstrap from Section 6 is restricted to nonparametric estimation procedures having slower rate of convergence than $1/\sqrt{n}$. Although designed as a regression bootstrap in the independent setup, it is consistent and hence robust against some form of dependence. Its advantage is simplicity, since no tuning parameter governing strength of general 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: the method may be outperformed by the block bootstrap.

8. OTHER RESULTS AND NOTES TO REFERENCES

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

Literature about the block bootstrap is extensive by now. A review of 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, are given in Radulović (1996a). Various results in empirical processes include Bühlmann (1994, 1995), Radulović (1996b, 1998) and Peligrad (1998). Lahiri (1996) 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 (cf. Politis and Romano, 1993). A version of the block bootstrap achieving stationarity for the bootstrap sample, the so-called stationary bootstrap, was proposed by Politis and Romano (1994a). Lahiri (1999) shows rigorously that the block bootstrap is better than the stationary bootstrap. Carlstein et al. (1998) propose a linking scheme for resampling blocks: they argue that, for the case of variance estimation of $\hat{\theta} = \overline{X}_n$, such a procedure has lower mean squared error. The tapered block bootstrap (Paparoditis and Politis, 2001) achieves this goal as well.

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

Model-based bootstrapping has been studied in numerous cases: Freedman (1984) and Bose (1988) for AR; Kreiss and Franke (1992) for ARMA; Rajarshi (1990), Paparoditis and Politis (2002) for Markov models. A nonparametric AR(1) model with heteroscedastic innovations is discussed in Franke, Kreiss, Mammen and Neumann (1998): this model-based bootstrap can be used for accurate construction of simultaneous confidence bands of the autoregression function $m(x) = \mathbb{E}[X_t \mid X_{t-1} = x]$. Note that the same can be achieved (to first order) by a local bootstrap from Section 6.

For the AR-sieve bootstrap, empirical process results are given in Bickel and Bühlmann (1999) via 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), where AR-sieve bootstrap confidence intervals for the trend are established.

Combining model- or sieve-based methods with the block bootstrap was suggested by Davison and Hinkley (1997, Chapter 8.2): they call the procedure *postblackening*. The idea is to prewhiten 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 postblackened resample.

Another way of bootstrapping stationary linear time series was proposed by Dahlhaus and Janas (1996): they independently resample periodogram values in the frequency domain according to a spectral density estimate. By construction, this resampling considers only the autocovariance structure and consistency is thus restricted to linear time series. The idea of independent resampling in the frequency domain appeared earlier 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 (1999).

ACKNOWLEDGMENTS

Comments by a reviewer and an Editor were particularly helpful. I also wish to thank Alexander McNeil for interesting remarks.

REFERENCES

ACCOLA, C. (1998). Bootstrap für die bedingte Erwartung bei Zeitreihen. Diploma thesis, ETH Zürich. (In German.)

ANGO NZE, P., BÜHLMANN, P. and DOUKHAN, P. (2002). Weak dependence beyond mixing and asymptotics for nonparametric regression. *Ann. Statist.* **30** To appear.

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

BICKEL, P. J. and BÜHLMANN, P. (1997). Closure of linear processes. *J. Theoret. Probab.* **10** 445–479.

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.

BOSE, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Ann. Statist.* **16** 1709–1722.

BRILLINGER, D. R. (1997). Random process methods and environmental data: The 1996 Hunter Lecture. *Environmetrics* **8** 269–281.

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

BÜHLMANN, P. (1995). The blockwise bootstrap for general empirical processes of stationary sequences. *Stochastic Process. Appl.* **58** 247–265.

BÜHLMANN, P. (1997). Sieve bootstrap for time series. *Bernoulli* **3** 123–148.

BÜHLMANN, P. (1998). Sieve bootstrap for smoothing in nonstationary time series. *Ann. Statist.* **26** 48–83.

BÜHLMANN, P. (2002). Sieve bootstrap with variable length Markov chains for stationary categorical time series. *J. Amer. Statist. Assoc.* To appear.

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

BÜHLMANN, P. and KÜNSCH, H. R. (1999). Block length selection in the bootstrap for time series. *Comput. Statist. Data Anal.* **31** 295–310.

BÜHLMANN, P. and WYNER, A. J. (1999). Variable length Markov chains. *Ann. Statist.* **27** 480–513.

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.

CARLSTEIN, E., DO, K.-A., HALL, P., HESTERBERG, T. and KÜNSCH, H. R. (1998). Matched-block bootstrap for dependent data. *Bernoulli* **4** 305–328.

CHOI, E. and HALL, P. (2000). Bootstrap confidence regions computed from autoregressions of arbitrary order. *J. Roy. Statist. Soc. Ser. B* **62** 461–477.

DAHLHAUS, R. and JANAS, D. (1996). A frequency domain bootstrap for ratio statistics in time series analysis. *Ann. Statist.* **24** 1934–1963.

DAVISON, A. C. and HALL, P. (1993). On Studentizing and blocking methods for implementing the bootstrap with dependent data. *Austral. J. Statist.* **35** 215–224.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge Univ. Press.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **82** 171–185.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

FRANKE, J. and HÄRDLE, W. (1992). On bootstrapping kernel spectral estimates. *Ann. Statist.* **20** 121–145.

FRANKE, J., KREISS, J.-P., MAMMEN, E. and NEUMANN, M. H. (1998). Properties of the nonparametric autoregressive bootstrap. Preprint, Univ. Kaiserslautern.

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

GÖTZE, F. and KÜNSCH, H. R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. *Ann. Statist.* **24** 1914–1933.

GRENANDER, U. (1981). *Abstract Inference.* Wiley, New York.

HALL, P. (1985). Resampling a coverage pattern. *Stochastic Process. Appl.* **20** 231–246.

HALL, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14** 1431–1452.

HALL, P. (1987). On the bootstrap and continuity correction. *J. Roy. Statist. Soc. Ser. B* **49** 82–89.

HALL, P., HOROWITZ, J. L. and JING, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika* **82** 561–574.

HALL, P., JING, B.-Y. and LAHIRI, S. N. (1998). On the sampling window method under long range dependence. *Statist. Sinica* **8** 1189–1204.

KREISS, J.-P. (1992). Bootstrap procedures for AR(∞)-processes. In *Bootstrapping and Related Techniques* (K.-H. Jöckel, G. Rothe and W. Sendler, eds.) 107–113. Springer, Berlin.

KREISS, J.-P. and FRANKE, J. (1992). Bootstrapping stationary autoregressive moving-average models. *J. Time Ser. Anal.* **13** 297–317.

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

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

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.

LAHIRI, S. N. (1996). On Edgeworth expansion and moving block bootstrap for Studentized M-estimators in multiple linear regression models. *J. Multivariate Anal.* **56** 42–59.

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

LÉGER, C., POLITIS, D. N. and ROMANO, J. P. (1992). Bootstrap technology and applications. *Technometrics* **34** 378–398.

LI, H. and MADDALA, G. S. (1996). Bootstrapping time series models. *Econometric Rev.* **15** 115–158.

LOH, W. (1987). Calibrating confidence coefficients. *J. Amer. Statist. Assoc.* **82** 155–162.

MARTIN, R. D. and YOHAI, V. J. (1986). Influence functionals for time series (with discussion). *Ann. Statist.* **14** 781–855.

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

NEUMANN, M. H. and KREISS, J.-P. (1998). Regression-type inference in nonparametric autoregression. *Ann. Statist.* **26** 1570–1613.

PAPARODITIS, E. and POLITIS, D. N. (1999). The local bootstrap for periodogram statistics. *J. Time Ser. Anal.* **20** 193–222.

PAPARODITIS, E. and POLITIS, D. N. (2000). The local bootstrap for kernel estimators under general dependence conditions. *Ann. Inst. Statist. Math.* **52** 139–159.

PAPARODITIS, E. and POLITIS, D. N. (2001). Tapered block bootstrap. *Biometrika* **88** 1105–1119.

PAPARODITIS, E. and POLITIS, D. N. (2002). The local bootstrap for Markov processes. *J. Statist. Plann. Inference.* To appear.

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

POLITIS, D. N. and ROMANO, J. P. (1992). A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectra density estimation. *Ann. Statist.* **20** 1985–2007.

POLITIS, D. N. and ROMANO, J. P. (1993). Nonparametric resampling for homogeneous strong mixing random fields. *J. Multivariate Anal.* **47** 301–328.

POLITIS, D. N. and ROMANO, J. P. (1994a). The stationary bootstrap. *J. Amer. Statist. Assoc.* **89** 1303–1313.

POLITIS, D. N. and ROMANO, J. P. (1994b). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** 2031–2050.

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling.* Springer, New York.

RADULOVIĆ, D. (1996a). The bootstrap of the mean for strong mixing sequences under minimal conditions. *Statist. Probab. Lett.* **28** 65–72.

RADULOVIĆ, D. (1996b). The bootstrap for empirical processes based on stationary observations. *Stochastic Process. Appl.* **65** 259–279.

RADULOVIĆ, D. (1998). The bootstrap of empirical processes for α-mixing sequences. In *High Dimensional Probability* (E. Eberlein, M. Hahn and M. Talagrand, eds.) 315–330. Birkhäuser, Basel.

RAJARSHI, M. B. (1990). Bootstrap in Markov-sequences based on estimates of transition density. *Ann. Inst. Statist. Math.* **42** 253–268.

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

ROBINSON, P. M. (1983). Nonparametric estimators for time series. *J. Time Series Anal.* **4** 185–207.

SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap.* Springer, New York.

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

WOODROOFE, M. and JHUN, M. (1989). Singh's theorem in the lattice case. *Statist. Probab. Lett.* **7** 201–205.