Bernoulli **4**(3), 1998, 305–328

Matched-block bootstrap for dependent data

EDWARD CARLSTEIN[1], KIM-ANH DO[2,3], PETER HALL[3], TIM HESTERBERG[3,4] and HANS R. KÜNSCH[3,5*]

[1]*Department of Statistics, University of North Carolina, Chapel Hill, NC 27514, USA*
[2]*Epidemiology and Population Health Unit, Queensland Institute of Medical Research, 300 Herston Road, Herston, Queensland 4006, Australia*
[3]*Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*
[4]*Research Department, MathSoft, 1700 Westlake Avenue, Suite 500, Seattle, WA 9109–3044, USA*
[5]*Seminar für Statistik, Eidgenössische Technische Hochschule, Zentrum, CH–8092 Zürich, Switzerland*

The block bootstrap for time series consists in randomly resampling blocks of consecutive values of the given data and aligning these blocks into a bootstrap sample. Here we suggest improving the performance of this method by aligning with higher likelihood those blocks which match at their ends. This is achieved by resampling the blocks according to a Markov chain whose transitions depend on the data. The matching algorithms that we propose take some of the dependence structure of the data into account. They are based on a kernel estimate of the conditional lag one distribution or on a fitted autoregression of small order. Numerical and theoretical analyses in the case of estimating the variance of the sample mean show that matching reduces bias and, perhaps unexpectedly, has relatively little effect on variance. Our theory extends to the case of smooth functions of a vector mean.

Keywords: blocking methods; bootstrap; kernel methods; resampling; time series; variance estimation

1. Introduction

In their classical form, as first proposed by Efron (1979), bootstrap methods are designed for application to samples of independent data. Under that assumption they implicitly produce an adaptive model for the marginal sampling distribution. During the last decade these approaches have been modified to suit the case of dependent data. Indeed, block bootstrap methods in that setting were introduced by Hall (1985), Carlstein (1986), Künsch (1989) and Liu and Singh (1992). They involve implicitly computing empirical models for the general multivariate distribution of a stationary time series, or even a more general data sequence, under particularly mild assumptions on the process generating the data. The models are of course highly adaptive, or nonparametric, in the spirit of bootstrap methods. Since the introduction of the blockwise bootstrap, the method has been investigated in quite some detail. Shao and Yu (1993), Naik-Nimbalkar and Rajarshi (1994), Bühlmann (1994, 1995), Radulovic (1995, 1996a, b), Politis and Romano (1992) and Bühlmann and Künsch (1995)

[*]To whom correspondence should be addressed. E-mail: kuensch@stat.math.ethz.ch

established consistency for a large number of statistics and processes generating the data. Distribution estimation by the block bootstrap has been studied by Lahiri (1991, 1996) and Götze and Künsch (1996), showing that the block bootstrap can produce second-order correct estimators, and by Davison and Hall (1993), pointing out the need to select carefully the variance estimator when using a percentile-t version of the block bootstrap. Hall and Jing (1996), Hall *et al*. (1995), and Bühlmann and Künsch (1994) have addressed the issue of block choice and related matters. Politis and Romano (1994, 1995) studied modifications of the basic procedure.

The block bootstrap relies on producing a compromise between preserving the dependence structure of the original data and corrupting it by supposing that the data are independent. Blocks of data are resampled randomly with replacement from the original time series, and then a simulated version of the original process is assembled by joining the blocks together in random order. Although blocks of data are dependent in the original time series, they are independent in the bootstrap version. This causes bias in the bootstrap variance which can be large if the dependence in the data is strong. It is to be hoped that performance could be improved by matching the blocks in some way, i.e. by using a block joining rule which in some sense favoured blocks that were a priori more likely to be close to one another. In the present paper we analyse this procedure both numerically and theoretically. We show that in an important class of situations it does indeed produce improved performance.

There is a variety of ways in which matching can be effected. In Section 2 we present a number of matching rules which adapt to some extent to the nature of the data, e.g. by assuming a Markovian dependence or an autoregressive model. However, since the analysis of the matched-blocks bootstrap is extremely difficult, we investigate mainly the case where blocks with similar values at the ends are paired. This is particularly appropriate when the data are generated by a continuous time process which is densely sampled so that the variance of the arithmetic mean decays at a slower rate than $O(n^{-1})$. Our results show that, in this context, simple matching rules produce variance estimators that are less biased than, and have virtually the same variability as, those based on the ordinary unmatched-block bootstrap. In the case of a Markov process the bias reduction is an order of magnitude, but in general the bias is reduced by a constant factor. The result on the variability is somewhat unexpected; one might have predicted that variance increases as a result of block matching, since it effectively introduces additional terms to the formula for the estimator. However, it turns out that the influence of those terms on variability is of second order.

Section 2 introduces a variety of matched-block bootstrap methods. Their asymptotic properties are sketched in Section 3. These results are supported by simulation experiments in Section 4 and by rigorous arguments in Section 5. This leads to the main conclusion of this paper, namely that the matched-block bootstrap enhances performance by reducing the effect of bias, with relatively little influence on variance.

We should mention here other methods which also reduce the bias. Künsch (1989) proposed a blockwise jackknife with smooth transition between observations left out and observations with full weight and similarly a weighted bootstrap. Politis and Romano (1995) suggested variance estimators that are essentially linear combinations of two block bootstrap estimators based on different block sizes.

2. Methodology for matching blocks

Given data $\mathscr{X} = \{X_i, 1 \leqslant i \leqslant n\}$ from a stationary time series, prepare blocks $\mathscr{B}_1, \ldots, \mathscr{B}_b$ where $\mathscr{B}_i = \{X_{i1}, \ldots, X_{il}\}$ is of length l. For overlapping blocks, $b = n - l + 1$ and $X_{ij} = X_{i+j-1}$. In the case of non-overlapping blocks we take b to be the integer part of n/l and $X_{ij} = X_{(i-1)l+j}$.

The matched-block bootstrap constructs a Markov chain on the blocks with transition probabilities depending on the data $\mathscr{X}$. Specifically, if $\mathscr{B}_{i_1}, \ldots, \mathscr{B}_{i_j}$ are the first j blocks, then the probability that the $(j+1)$th block is $\mathscr{B}_{i_{j+1}}$ equals $p(i_j, i_{j+1})$. The first block is chosen according to the stationary distribution of the chain. As we shall see below, for our choices of the transition probabilities the stationary distribution is close to the uniform. So we can start the chain also with the uniform distribution. The blocks obtained in this way are then put into a string $\mathscr{B}_{i_1}, \mathscr{B}_{i_2}, \ldots$. The first n values of this string constitute the bootstrap resample $\mathscr{X}^*$. If T is a function of n variables (representing the data) and $\hat{\theta} = T(\mathscr{X})$ is an estimator of an unknown parameter θ, then generally $\hat{\theta}^* = T(\mathscr{X}^*)$ is its bootstrap version. The percentile form of the bootstrap estimates $\mathrm{var}(\hat{\theta})$ by $\mathrm{var}'(\hat{\theta}^*)$ and $P(\hat{\theta} - \theta \leqslant t)$ by $P'(\hat{\theta}^* - \hat{\theta} \leqslant t)$ where the prime denotes conditioning on the data $\mathscr{X}$. Centrings other than $\hat{\theta}$ are possible. For example, if $\overline{X}$ denotes the sample mean, then $\mathrm{E}'(\overline{X}^*) \neq \overline{X}$ because the stationary distribution of the Markov chain will generally not be exactly uniform on the blocks, and not all observations appear in an equal number of blocks. However, the latter effect is only a boundary one, and the stationary distribution is in general quite close to being uniform.

Construction of the transition probabilities $p(i_1, i_2)$ is the essential part of our algorithm. Ideally we would do it in such a way that the bootstrap samples have properties similar to those of the original sample. On the other hand, there should be sufficient variability to produce a rich class of simulations, rather than simply reproducing the original sample. Our proposals achieve this by matching the blocks only through their values near the beginnings or ends of blocks. The simplest proposal is *kernel matching*, where (for non-overlapping blocks)

$$p(i_1, i_2) \propto \begin{cases} K\left(\dfrac{X_{i_1,l} - X_{i_2-1,l}}{h}\right) & \text{if } i_2 \neq 1, \\ K\left(\dfrac{X_{i_1+1,1} - X_{11}}{h}\right) & \text{if } i_2 = 1,\ i_1 l < n, \\ 0 & \text{if } i_2 = 1,\ i_1 l = n. \end{cases} \tag{2.1}$$

Here, K is a symmetric probability density and h is a bandwidth. The proportionality constant for each i_1 is determined by the requirement that, for all i_1, $\sum_{i_2} p(i_1, i_2) = 1$.

Note that we match the last observation in $\mathscr{B}_{i_1}$ with the last observation in the block preceding $\mathscr{B}_{i_2}$ in the original sample. Implicit in the matching rule is an assumption that the dependence is mainly of Markovian character, since we use only the last observation in the block $\mathscr{B}_{i_1}$ to determine where $\mathscr{B}_{i_2}$ should start. In other words, the matching rule (2.1) corresponds to choosing the first element of $\mathscr{B}_{i_2}$, conditional on the last element of $\mathscr{B}_{i_1}$, according to the conditional distribution of X_i given X_{i-1}.

Alternatively, we can replace the observations X_i by their ranks. We call this *rank matching*. Our basic rank-matching method produces a stationary distribution which is exactly uniform. We let $R_i^{(\text{end})}$ be the rank of X_{il} among $X_{1l}, \ldots, X_{bl}$, and $R_i^{(\text{start})}$ be the rank of X_{il} among $X_{0l}, \ldots, X_{(b-1)l}$, with $X_{0l} = X_{1l}$ and ties broken arbitrarily. Now letting 1_A denote the indicator function of a set A, we set $p(j_1, j_2) = q(R_{j_1}^{(\text{end})}, R_{j_2-1}^{(\text{start})})$, where

$$q(i, j) = (2m+1)^{-1}(1_{\{|i-j| \leqslant m\}} + 1_{\{i+j \leqslant m+1\}} + 1_{\{i+j \geqslant 2b+1-m\}}).$$

This defines a doubly stochastic matrix, i.e. one where all row and column sums are equal to one. We also considered a modified rank-matching procedure that is roughly equivalent to kernel matching with actual values replaced by normal scores, but implemented by a method that requires time $O(1)$ rather than $O(b)$ for computing each transition.

Obviously we can also extend kernel and rank matching to the case where more than one observation (at the end of block i_1 and the block preceding i_2) is used for the matching, in particular by taking products of kernels. This very quickly becomes impractical, however, because of the curse of dimensionality; either $p(i_1, i_2)$ is almost constant (if the bandwidth is large) or $p(i_1, i_1 + 1)$ dominates (if the bandwidth is small). An alternative procedure, *autoregressive matching*, takes into account $p < l$ observations at the ends of blocks. It is based on a fitted AR(p) model, with coefficients $\hat{\phi}_1, \ldots, \hat{\phi}_p$ and distribution of the innovations given by $\hat{F}_\epsilon$. By iterating the defining equation of the model we produce matrices $A(\hat{\phi})$ and $B(\hat{\phi})$ such that

$$U_{i+p} = A(\hat{\phi})U_i + B(\hat{\phi})(\epsilon_{i+p}, \ldots, \epsilon_{i+1})',$$

where $U_i = (X_i, \ldots, X_{i-p+1})'$. This suggests the following algorithm. If the current block is $\mathscr{B}_{i_1}$, generate $\epsilon_1^*, \ldots, \epsilon_p^*$ by sampling independently from $\hat{F}_\epsilon$, and take the next block to be $\mathscr{B}_{i_2}$, where i_2 minimizes the L_1 norm of

$$(\epsilon_p^*, \ldots, \epsilon_1^*)' - B(\hat{\phi})^{-1}\{U_{(j-1)l+p} - A(\hat{\phi})U_{i_1 l}\}$$

over j. This amounts to choosing the first p values of the next block according to the fitted model, up to a discretization error. Autoregressive matching is thus a compromise between the AR bootstrap (Efron and Tibshirani 1993) and the independent block bootstrap. Further details are given in the fourth paragraph of Section 3. Other ways to match are possible; we could for instance match that linear combination of values at the end of the blocks which predicts the average of future values best. For statistics other than mean, and for multivariate time series, we could match based on block ends or on linear combinations of values of a univariate time series obtained by replacing the (multivariate) observations with a measure of the influence of individual observations, such as jackknife values. These topics are left open for future research.

Empirical choice of block length may be achieved by modifying methods suggested by Hall and Jing (1996). We outline the argument here. Observe first that the variance of a block bootstrap estimator of variance, with or without block matching, is generally asymptotic to a constant multiple of l/n^3, and that the squared bias is asymptotic to a constant multiple of $(nl)^{-2}$. (When there is no matching this result is due to Hall (1985), Carlstein (1986) and Künsch (1989). With matching, and in the non-Markovian case, it is derived in Section 5; see for example Remark 1.) Therefore, the asymptotically optimal block length is $l_n = Cn^{1/3}$, where $C > 0$ depends on characteristics of the time series and the blocking method. Essentially, the problem of block choice is one of estimating C.

Suppose that we have an estimator $\hat{l}_m$ of l_m, for $m < n$. Then $\hat{l}_n = (m/n)^{1/3}\hat{l}_m$ is an estimator of l_n. We may construct $\hat{l}_m$ as follows. There are $k = n - m + 1$ subseries of length m that may be obtained from the full time series of length n. Let $\hat{\theta}_n$ denote the statistic of interest computed for the full series, let $\hat{\theta}_{mi}$ (for $1 \leqslant i \leqslant k$) be the statistics based on the subseries, and put $\sigma^2_m = \mathrm{var}(\hat{\theta}_{mi})$. An estimator of σ^2_m is $\hat{\sigma}^2_m = k^{-1}\sum_i(\hat{\theta}_{mi} - \hat{\theta}_n)^2$. Let $\tilde{\sigma}^2_{mi}(l)$ be the estimator of σ^2_m computed using a given block bootstrap method applied to the ith subseries and employing block length l. An estimator of the mean squared error $s_m(l) = \mathrm{E}\{\tilde{\sigma}^2_{mi}(l) - \sigma^2_m\}^2$ is

$$S_m(l) = k^{-1}\sum_{i=1}^{k}\{\tilde{\sigma}^2_{mi}(l) - \hat{\sigma}^2_m\}^2.$$

We may choose $\hat{l}_m$ to minimize $S_m(l)$, and put $\hat{l}_n = (m/n)^{1/3}\hat{l}_m$.

It is straightforward to prove that $\hat{l}_n$ is consistent for l_n, in the sense that $\hat{l}_n/l_n \to 1$ in probability, if $m = o(n^{3/16})$. (Note that $S_m(l) = s_m(l) + O_p(n^{-1/2})$, $s_m(l)$ is of size $m^{-8/3}$ if l is of the optimal size $n^{1/3}$, and $m^{-8/3}$ is of larger order than $n^{-1/2}$ if $m = o(n^{3/16})$.) A longer argument shows that substantially larger orders of m also produce consistent results. In practice we shall have to choose m in advance, but the effect of this is of second order and thus less crucial than that of choosing l which is of first order. Establishing the numerical performance of empirical choice of block size is beyond the scope of this paper, however.

3. Overview of large-sample properties

We consider first the problem of estimating the variance of the sample mean using non-overlapping blocks, and then discuss more general statistics and overlapping blocks. For independent non-overlapping blocks we have

$$\mathrm{E}\{\mathrm{var}'(\overline{X}^*)\} - \mathrm{var}(\overline{X}) \sim -\beta_1 = -2(nl)^{-1}\sum_{j=1}^{\infty} j\,\mathrm{cov}(X_0, X_j) \tag{3.1}$$

and

$$\mathrm{var}\{\mathrm{var}'(\overline{X}^*)\} \sim 2n^{-3}l\,\mathrm{var}(\overline{X})^2. \tag{3.2}$$

We argue that, for a wide range of matching rules, (3.2) remains the same, but

$$\mathrm{E}\{\mathrm{var}'(\overline{X}^*)\} - \mathrm{var}(\overline{X}) \sim -\beta_1 + \beta_2, \tag{3.3}$$

where β_2 is generally of the same sign as β_1 (for a non-repulsive process). In other words, block matching changes (and often reduces) the bias but has relatively little effect on variance.

These properties will be derived rigorously in Section 5, for a slightly simplified procedure and a specific class of matching rules. We give here a simple recipe for calculating β_2 for general matching rules. Then we apply it to the rules introduced in Section 2.

The first step is to simplify the formula for the transition probabilities $p(i_1, i_2)$. We suppose that, to a first approximation,

$$p(i_1, i_2) \sim b^{-1}\phi(U_{i_1}, V_{i_2-1}), \tag{3.4}$$

where U_i and V_i are functions of (X_{ij}) for j close to l. This reflects the fact that matching occurs mainly through the values near the ends of the blocks. The property $\sum_{i_2} p(i_1, i_2) = 1$ translates to

$$\mathrm{E}\{\phi(u, V_i)\} \equiv 1. \tag{3.5}$$

For (3.2) to hold we need the stationary distribution of the chain to be approximately uniform. This means that

$$\mathrm{E}\{\phi(U_i, v)\} \equiv 1. \tag{3.6}$$

Finally, the formula for β_2 is

$$\beta_2 = 2(nl)^{-1} \sum_{i=-\infty}^{0} \sum_{k=1}^{\infty} \mathrm{E}\{\mathrm{E}(X_i - \mu | U_0)\mathrm{E}(X'_k - \mu | V'_0)\phi(U_0, V'_0)\}, \tag{3.7}$$

where $\mu = \mathrm{E}(X_i)$ and $\{X'_i\}$ is an independent copy of $\{X_i\}$ (and V'_i is defined in terms of X'_j). In Section 5 it will become clear why this formula is to be expected.

Let us compute (3.4) and (3.7) for the matching rules of Section 2. For kernel matching we assume that X_i has density f. The law of large numbers suggests that the proportionality constant in (2.1) is

$$b\int K\left(\frac{X_{i_1 l} - y}{h}\right) f(y)\,\mathrm{d}y.$$

Letting the bandwidth h tend to zero we obtain formally (3.4), with $U_i = V_i = X_{il}$ and $\phi(u, v) = f(u)^{-1}\delta(u - v)$, where δ denotes the Dirac delta function. Note that (3.5) and (3.6) are satisfied. Moreover,

$$\beta_2 = 2(nl)^{-1} \sum_{i=-\infty}^{0} \sum_{k=1}^{\infty} \mathrm{E}\{\mathrm{E}(X_i - \mu | X_0)\mathrm{E}(X_k - \mu | X_0)\}. \tag{3.8}$$

For Gaussian processes, (3.8) can be expressed with the covariance function. Moreover, if $\{X_i\}$ is a Markov process, then $\{X_i, i < 0\}$ and $\{X_k, k > 0\}$ are conditionally independent, given X_0. Thus,

$$\mathrm{E}\{\mathrm{E}(X_i - \mu | X_0)\mathrm{E}(X_k - \mu | X_0)\} = \mathrm{E}[\mathrm{E}\{(X_i - \mu)(X_k - \mu) | X_0\}] = \mathrm{cov}\,(X_i, X_k),$$

whence

$$\beta_2 = 2(nl)^{-1} \sum_{i=-\infty}^{0} \sum_{k=1}^{\infty} \mathrm{cov}\,(X_i, X_k) = 2(nl)^{-1} \sum_{j=1}^{\infty} j\,\mathrm{cov}\,(X_0, X_j) = \beta_1.$$

This result and (3.3) show that, for Markov processes, kernel matching reduces the bias of the bootstrap variance by an order of magnitude. That is understandable, since kernel matching relies on a Markovian assumption.

With rank matching the results are the same as for kernel matching. So we turn to autoregressive matching. There we assume that the process $\{X_i\}$ is AR(p), that the innovations ϵ_i have a density g_1, and that the estimators $\hat{\phi}_j$ and $\hat{F}_\epsilon$ are consistent. Set $U_i = (X_{il}, \ldots, X_{i,l-p+1})'$, $V_i = (X_{i+1,p}, \ldots, X_{i+1,1})'$ and

$$g_p(x_1, \ldots, x_p) = \prod_{i=1}^{p} g_1(x_i).$$

Denote the density of U_i by f. Then we expect (3.4) to hold with

$$\phi(u, v) = f(v)^{-1} g_p[B(\phi)^{-1}\{v - A(\phi)u\}].$$

Because $f(u)g_p[B(\phi)^{-1}\{v - A(\phi)u\}]$ is the joint density of (U_i, V_i) then (3.5) and (3.6) are satisfied, and (3.7) becomes

$$\beta_2 = 2(nl)^{-1} \sum_{i=-\infty}^{0} \sum_{k=1}^{\infty} \mathrm{E}\{\mathrm{E}(X_i - \mu | U_0)\mathrm{E}(X_k - \mu | V_0)\}.$$

An AR(p) process is Markovian of order p; so we obtain by the same argument as before that again $\beta_2 = \beta_1$. Therefore, autoregressive matching also reduces bias by an order of magnitude, provided that the model behind the matching rule is correct.

The case of a nonlinear statistic, $\hat{\theta}$, is in principle similar, yielding the same results. That is, variance is unaffected to first order by block matching (for non-overlapping blocks), but bias can be reduced by a constant factor, or by an order of magnitude if the time series is Markovian. Theoretical arguments are as follows, in outline.

Assume that $\hat{\theta}$ can be written as a smooth function of a vector of means of functions of the time series data values. Taylor-expand this quantity about the expected values of the means, producing terms T_0 (constant), T_1 (linear), T_2 (quadratic), and so on: $\hat{\theta} = T_0 + T_1 + T_2 + \ldots$. It may be proved that $\mathrm{var}(\hat{\theta}) = \mathrm{var}(T_1) + O(n^{-2})$. The block bootstrap approximation to variance, with or without matching, may be decomposed in the same way. Since its error in (implicitly) approximating $\mathrm{var}(T_1)$ is of strictly larger size than n^{-2} (compare for example (3.2)), then the implicit bootstrap approximation to the linear component dominates. As a result, the same properties are exhibited by the full approximation, to first order.

It is not difficult to see that the bias reduction property holds also for the more complex overlapping-blocks method, although the variance reduction result is more difficult to verify rigorously. In our detailed theoretical analysis (Section 5) we shall confine attention to non-overlapping blocks, but our numerical work (Section 4) will confirm that block matching improves overall performance for both overlapping and non-overlapping blocks. The improvement provided by matching non-overlapping blocks is significantly greater than that offered by allowing non-overlapping blocks to overlap, although blocks that are *both* matched and overlapping perform best of all.

We do not have a satisfactory theoretical account of the matched-block bootstrap for distribution estimation and so do not examine it in the present paper. It is possible to argue that the matched-block bootstrap is consistent in the sense that $P'[n^{1/2}(\overline{X}^* - \overline{X}) \leqslant x]$ converges almost surely to the correct asymptotic normal probability, but the real issue is whether it *improves* over the normal approximation as is the case for the unmatched block bootstrap (Lahiri 1991, 1996, Götze and Künsch 1996). It is likely that similar results hold also for the matched-block bootstrap.

4. Numerical results

Our simulation study focused on the estimated variance of the sample mean. We measured the accuracy using the mean squared error (MSE) of the logarithm of the variance (using the MSE for the variance would favour estimates which are biased downwards owing to a shrinkage effect). We employed two models.

Model 1 (*first-order autoregressive processes*) is defined by $X_{t+1} = \rho X_t + (1-\rho^2)^{1/2}\epsilon_{t+1}$ for $\rho = 0.95$ and $\rho = 0.80$. The independent innovations ϵ_t were normal $N(0, 1)$.

Model 2 (*exponentially decaying covariance*) has X as stationary Gaussian processes with covariance function $\gamma(t) = \operatorname{cov}(X_s, X_{s+t}) = \exp(-c|t|^\alpha)$, where $c = 0.000\,15$ and $\alpha = 1.5$ or 1.95. (Model 1 addresses the case $\alpha = 1$.) We simulated these processes using the algorithm developed by Wood and Chan (1994).

Sample sizes ranged from 200 to 5000. We did not consider each possible combination of parameters, statistics, models and block matching and devoted greatest attention to Model 1, which is the case on which we report below. However, broadly similar results were obtained in all cases.

For Model 1 and in the context of non-overlapping blocks, we considered six different matching methods:

(1) *unmatched:* block bootstrap with no matching;

(2) *uniform kernel:* the method suggested by (2.1), with initial probabilities taken to be uniform on blocks;

(3) *stationary kernel:* as for the uniform kernel, but in the stationary distribution;

(4) *modified kernel:* as for the uniform kernel, but with each column of the transition matrix multiplied by a constant so that the stationary distribution is indeed uniform;

(5) *rank:* the basic rank method suggested in Section 2;

(6) *modified rank:* the modified rank method suggested in Section 2.

For overlapping blocks we considered only the unmatched, rank and modified rank methods. Kernel methods could be used for overlapping blocks, but rank methods are preferable. In the case of overlapping blocks a penalty needs to be imposed on kernel rules, to counteract a strong tendency to match only neighbours or near neighbours.

We used the standard normal kernel for methods (2)–(4). (In the "uniform kernel" method, the qualifier "uniform" refers to the initial distribution of block probabilities, not to the type of kernel.) The bandwidth was chosen as $h = cb^{-1/5}\hat{\sigma}$, where $\hat{\sigma}$ was the sample standard deviation of the simulated sequence, and $c = \frac{1}{2}$, 1 or 2. In the case $c = 1$, and except for the fact that the constant is 1 rather than 1.06, this is the "equivalent normal density" prescription for bandwidth selection (Silverman 1986, p. 45). For methods (5) and (6) we used $h = cb^{-1/5}/[1 + (cb^{-1/5})^2]^{1/2}$, and for method (5) we took $m = 0.84bh$.

Our main conclusions are as follows.

(a) Block matching can substantially reduce the MSE, and in fact the MSE is consistently smaller using any one of the matching rules, relative to not using it. Since the main effect of matching is bias reduction, and not reduction of variance, optimal performance for matched-block rules is obtained using relatively small blocks. As predicted by our theory, the variance component of the MSE changes little as a result of matching.

(b) The unmatched and rank methods were very fast to implement. At the other extreme, the modified kernel method was very slow, since it required each transition matrix to be stored and iteratively modified to be doubly stochastic. The kernel methods require time $O(b)$ for computing each transition. (This may be reduced by using a kernel with finite support, or by storing transition matrices.) The unmatched and rank methods need $O(1)$ time for each transition, the modified rank method being slower because it requires more floating-point operations. These computational issues would be particularly significant if resampling methods, such as those suggested in Section 2, were used to choose block length.

(c) Among the different matching methods, and for either overlapping blocks or non-overlapping blocks, the rank method generally gives least MSE, although the two rank methods are not far apart. Performance at the optimum for overlapping blocks matched using the rank method is generally superior to that of non-overlapping blocks using the rank method.

(d) The stationary kernel method gives the worst performance among the matching methods but is still better at the optimum (often a great deal better) than the unmatched method at its optimum. Among the kernel methods, the modified kernel approach usually produces the smallest MSE, and the uniform kernel method is second.

(e) The preferred choice overall, on grounds of both MSE performance and computational ease, is the unmodified rank method. It might be improved by replacing the single uniform random number U that drives each transition with, for example, $(U_1 + U_2 + U_3 + U_4)/2$, in effect replacing a uniform kernel with a more traditional symmetric unimodal kernel.

We economize on space by illustrating here only two parameter settings: $\rho = 0.8$ and $\rho = 0.95$ in Model 1, both for $n = 200$ and $b = 1$ (in the definition of bandwidth), for estimating the logarithm of the variance of the sample mean, and for unmatched, uniform kernel and rank methods respectively. In the cases of unmatched and rank methods, root mean squared error (RMSE) for both overlapping and non-overlapping blocks are illustrated, but only non-overlapping blocks are depicted in the case of the uniform kernel method (see Figure 1 ($\rho = 0.8$) and Figure 2 ($\rho = 0.95$)). These results show that, for the specific parameter settings illustrated, the methods at their optimal settings perform in the following order (from best to worst, in terms of MSE): rank matching with overlapping blocks, rank with non-overlapping, uniform kernel with non-overlapping, unmatched with overlapping, and unmatched with non-overlapping.

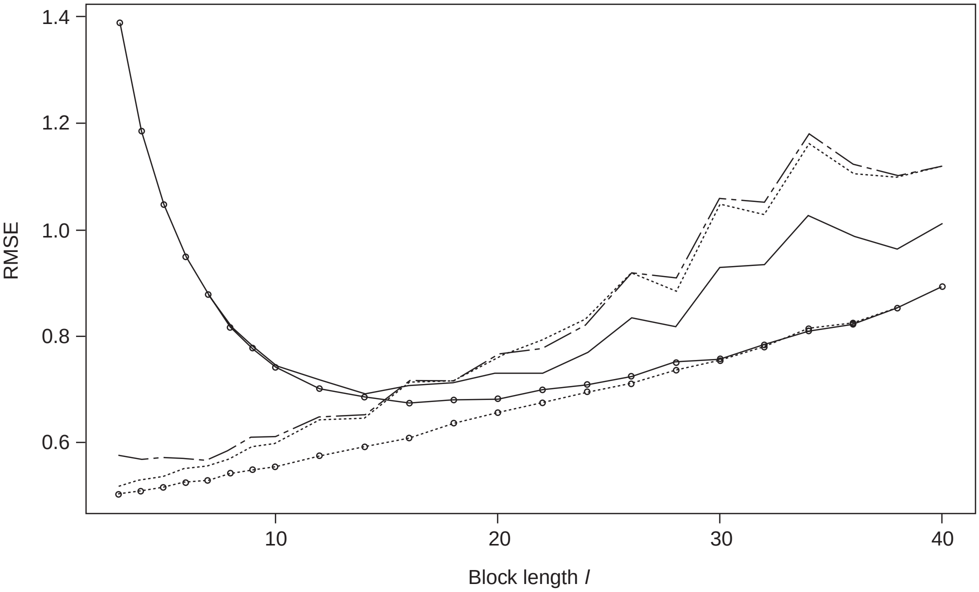

Fig. 1. RMSE in the case of relatively short-range dependence. RMSE (vertical axis) for estimating the logarithm of the variance of the sample mean, is plotted as a function of block length l (horizontal axis), for different block bootstrap estimators ((———), unmatched; (○——○), unmatched with overlap; (——–—), uniform kernel; (– – – –), rank; (○– – –○), rank with overlap), in the case of Model 1 with $\rho = 0.8$ and $n = 200$.

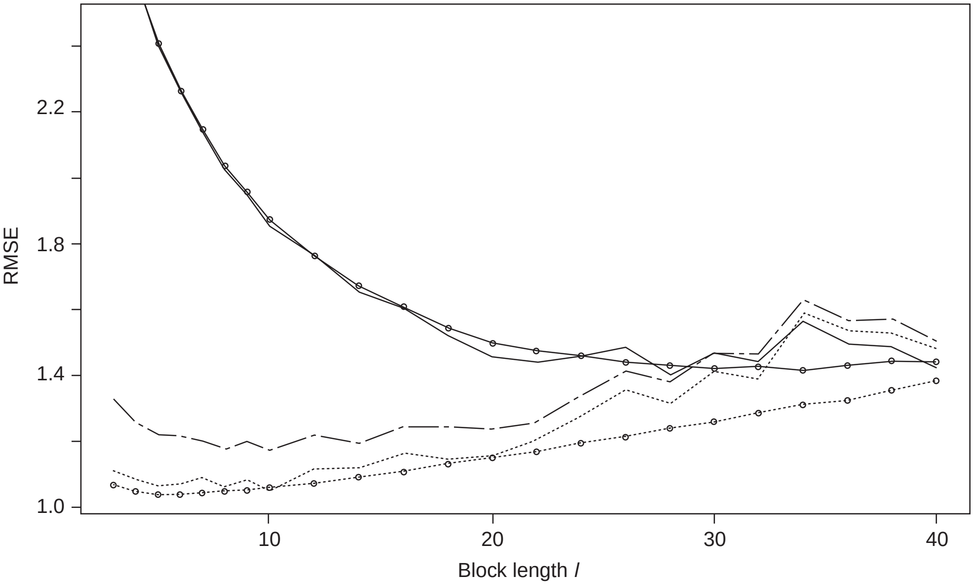

Fig. 2. RMSE in the case of relatively long-range dependence. Same as Figure 2, but with $\rho = 0.95$.

We conclude by summarizing numerical results that are not explicitly illustrated here. For optimal choice of block length, block matching substantially reduces bias in all cases considered, and use of overlapping blocks consistently reduces variance. However, the reduction in bias using block matching is greater, in proportionate terms, than that in variance using overlapping blocks. Hence, the improvement in overall performance using overlapping blocks is relatively small, when a good matching rule is employed. This is particularly true in the case of longer-range dependence (e.g. $\rho = 0.95$ in Model 1).

Also under longer-range dependence, the MSE depends relatively little on the block length; the MSE curve is shallower, and block lengths distant some way from the optimum produce virtually optimal performance. This implies that empirical methods for selecting block length would be likely to exhibit increased variability as the range of dependence increases. The shortness of the blocks which give optimal performance for matched rules, relative to those for unmatched rules, is substantially more pronounced in the case of longer-range dependence.

The performance of the kernel methods at their optimum block lengths is generally slightly improved by undersmoothing (i.e. multiplying the bandwidth by $c = \frac{1}{2}$), although at relatively large suboptimal block lengths it is slightly improved by oversmoothing (i.e. using $c = 2$). As expected, increasing the bandwidth tends to increase the bias component of MSE and to decrease the variance component. Choice of a symmetric non-negative kernel has substantially less effect on performance than does choice of bandwidth, provided that the kernel is rescaled so that it represents a distribution with unit variance. (We experimented with the Epanechnikov kernel.) There will be problems using high-order kernels, however, since then the transition probabilities can be expected to take negative values with positive probability.

All these characteristics become more marked as sample size increases. However, the differences between the different modifications of rules (the three different kernel rules, and the two different rank rules) decrease, in relative terms, as n increases. This is particularly true at the optimum for these rules, where for $n = 1000$ the performances of the three kernel rules are virtually indistinguishable (for non-overlapping blocks), as too are those of the two rank rules (for either overlapping or non-overlapping blocks).

5. Theoretical results

We now turn to rigorous derivation of results (3.2) and (3.3). The technical details of theory for block matching are particularly arduous. To keep them in manageable succinct form we treat a somewhat abstract version of the procedure that we discussed earlier in Sections 2–4. For simplicity we assume that $n = bl$ for integers b (the number of blocks) and l (the length of each block). We consider the case where time series data $\{X_i\}$ are derived by sampling a continuous process with a sampling frequency which may increase with n. So, let $\{Y(t),\ t \in (0, \infty)\}$ denote a stationary stochastic process in continuous time, implying that $\mu \equiv \mathrm{E}\{Y(t)\}$ does not depend on t, and $\gamma(t) \equiv \operatorname{cov}\{Y(s), Y(s+t)\}$ does not depend on s. Let $\lambda = \lambda(n)$ represent a sequence of positive constants possibly diverging to infinity as n increases, and put $X_i = Y(i/\lambda)$. The strength of dependence of the process $\{X_i\}$ increases with increasing λ. Indeed, the variance of the sample mean is of order $O(\lambda/n)$ (see below), and so long-range dependence might be considered to be characterized by the case where λ increases with sample size.

Our assumptions on the process Y are as follows.

Condition (C_{1A}). *Y is t_0 dependent for some $t_0 > 0$, meaning that the sigma-fields $\mathscr{F}(0, s)$ and $\mathscr{F}(s + t_0, \infty)$ generated by $\{Y(u), u \in (0, s)\}$ and $\{Y(u), u \in (s + t_0, \infty)\}$, respectively, are independent for each $s > 0$.*

Condition (C_{1B}). $\mathrm{E}|Y(t)|^{\alpha} < \infty$ *for some $\alpha > 8$, to be determined later.*

Condition (C_{1A}) simplifies our arguments, but modified versions of our results hold in the case where Y is mixing with geometrically decreasing mixing rate.

Next we set down our assumptions on the block-matching algorithm. Remember that $p(j_1, j_2)$ for $1 \leqslant j_1, j_2 \leqslant b$ is the data-dependent probability that the next block is $\mathscr{B}_{j_2}$, given that the current block is $\mathscr{B}_{j_1}$. Put $V_i = (X_{i1}, \ldots, X_{ir})$ and $U_i = (X_{i,l-r+1}, \ldots, X_{il})$, the first r and the last r values in block i respectively. We impose the following conditions.

Condition (C_{2A}). *For all j_1, j_2,*

$$p(j_1, j_2) = \psi(U_{j_1}, V_{j_2}; V_j, j \neq j_2),$$

where ψ is non-negative and symmetric in the last $b - 1$ arguments, and $r = O(\lambda)$.

Condition (C_{2B}). *For all j_1,*

$$\sum_{j_2=1}^{b} p(j_1, j_2) = 1.$$

Condition (C_{2C}). *For some $\epsilon > 0$,*

$$\sup_{j} \mathrm{E}\{p(j, j+1)\} = O(b^{-\epsilon}).$$

Condition (C_{2D}). *For any j_1, j_2, j_3, j_4 with $j_3 \neq j_2$, $j_4 \neq j_2$, there exists $p'(j_1, j_2, j_3, j_4) \in [0, 1]$ which depends only on U_{j_1} and V_j, $j \neq j_3, j_4$ such that for some $\epsilon > 0$ and all $1 \leqslant q < \infty$,*

$$\sup_{j_1, j_2, j_3, j_4} \|p(j_1, j_2) - p'(j_1, j_2, j_3, j_4)\|_q = O(b^{-1-\epsilon}),$$

where $\|.\|_q$ denotes the L_q norm.

Condition (C_{2E}). *For some $\epsilon > 0$,*

$$\operatorname{ess\,sup}_{j_2 \neq j_1+1} \mathrm{E}\{p(j_1, j_2) | V_{j_1}; \mathscr{B}_j, j \neq j_1, j_1 + 1\} = O(b^{-\epsilon}).$$

For example, we might define

$$
\begin{aligned}
p_1(j_1, j_2) &= \prod_{k=0}^{r-1} K\left(\frac{X_{j_1,l-k} - X_{j_2,r-k}}{h}\right), \\
p(j_1, j_2) &= p_1(j_1, j_2) \Bigg/ \left(\sum_{j=1}^{b} p_1(j_1, j)\right),
\end{aligned}
\tag{5.1}
$$

where $K \geqslant 0$ denotes a bounded and compactly supported kernel function and h is a bandwidth satisfying $h = O(b^{-\epsilon})$ and $b^{1-\epsilon}h \to \infty$ for some $\epsilon > 0$; and $1 \leqslant r = r(n) = o(\lambda)$. In the event that the denominator in the definition of $p(j_1, j_2)$ vanishes, define $p(j_1, j_2)$ to equal b^{-1} for each value of j_2. Conditions (C_2) may be verified in this setting, for a wide variety of processes including polynomial functions of Gaussian processes whose covariance γ satisfies Condition (C_{1A}). In this setting the approximating probability $p'(j_1, j_2, j_3, j_4)$ in Condition (C_{2D}) may be constructed by removing from the denominator in the definition of $p(j_1, j_2)$ a finite number of terms $p_1(j_1, j)$, so as to achieve the desired independence. Furthermore, Condition (C_{2E}) is an immediate consequence of the compact support of K and of the conditions imposed on h. Note that by way of contrast to rule (2.1), rule (5.1) now assumes strong positive dependence for neighbouring values, which is natural in the context of dense sampling of a continuous process.

Of course, many alternative prescriptions of p are possible, still satisfying Conditions (C_2). In particular, there is considerable latitude for varying the block representatives that are compared via the kernel function in the definition of p_1 at (5.1).

Let $\overline{X}$ and $\overline{X}^*$ denote sample means of the data $\mathscr{X}$ and resampled $\mathscr{X}^*$, respectively, and let

$$
\sigma^2 = \sigma^2(n) = \operatorname{var}(\overline{X}) = n^{-1}\left\{\gamma(0) + 2\sum_{j=1}^{n-1}(1 - n^{-1}j)\gamma\left(\frac{j}{\lambda}\right)\right\}
$$

represent the variance of the sample mean. The matched-block bootstrap estimator of σ^2 is given by $\hat{\sigma}^2 = \operatorname{var}'(\overline{X}^*)$, where the prime denotes conditioning on $\mathscr{X}$. To appreciate the size of the quantity that we are estimating, note that, if $\lambda \to L$ as $n \to \infty$, where $0 < L \leqslant \infty$, then

$$
\sigma^2 \sim \begin{cases} n^{-1}\left\{\gamma(0) + 2\displaystyle\sum_{j=0}^{\infty}\gamma\left(\frac{j}{L}\right)\right\} & \text{if } L < \infty, \\ 2\dfrac{\lambda}{n}\displaystyle\int_0^{\infty}\gamma(t)\,\mathrm{d}t & \text{if } L = \infty. \end{cases}
$$

Therefore, σ^2 is of size λ/n.

Let the stationary distribution on the block indices $(1, \ldots, b)$ be $\pi_1, \ldots, \pi_b$. Assume that the blocks $\mathscr{B}_{i_j}$ are produced with the chain in this stationary state, and put $\overline{X}' = \sum\pi_i\overline{X}_i$, $\overline{X}_i = l^{-1}\sum_j X_{ij}$. Then $\mathrm{E}'(\overline{X}^*) = \overline{X}'$ and $\hat{\sigma}^2 = b^{-2}(S_1 + 2S_2)$, where

$$S_1 = b\sum_{i=1}^{b} \pi_i(\overline{X}_i - \overline{X}')^2,$$

$$\begin{aligned}S_2 &= \sum_{j=1}^{b-1}\sum_{i=1}^{b-j} \mathrm{E}'\{(\overline{X}_j^* - \overline{X}')(\overline{X}_{i+j}^* - \overline{X}')\} \\ &= \sum_{i=1}^{b-1}(b-i)\mathrm{E}'\{(\overline{X}_1^* - \overline{X}')(\overline{X}_{i+1}^* - \overline{X}')\} \\ &= \sum_{i=1}^{b-1}(b-i)\sum_{j_1=1}^{b}\sum_{j_2=1}^{b} \pi_{j_1} p(i;\, j_1, j_2)(\overline{X}_{j_1} - \overline{X}')(\overline{X}_{j_2} - \overline{X}')\end{aligned}$$

and $p(i;\, j_1, j_2)$ denotes the i-step transition probability in the Markov chain of blocks $(p(1;\, j_1, j_2) = p(j_1, j_2))$.

If the stationary distribution of the block-matching rule is approximately uniform and is reached after two steps, then to a good approximation, $S_j \approx T_j$ where

$$T_1 = \sum_{i=1}^{b}(\overline{X}_i - \overline{X})^2, \qquad T_2 = \sum_{j_1=1}^{b}\sum_{j_2=1}^{b} p(j_1, j_2)(\overline{X}_{j_1} - \overline{X})(\overline{X}_{j_2} - \overline{X}).$$

This suggests an alternative variance estimator,

$$\tilde{\sigma}^2 = b^{-2}(T_1 + 2T_2).$$

We shall describe the theory for this quantity. We believe that $\tilde{\sigma}^2$ contains the essential features of $\hat{\sigma}^2$, for the following reasons. We showed earlier that the stationary distribution is uniform for a version of rank matching, and that it is approximately uniform in other cases since (3.6) is satisfied. That the stationary distribution is reached after two steps is plausible because V_{j_2} and U_{j_2} are independent if l is large. Hence, the two terms $p(j_1, j_2)$ and $p(j_2, j_3)$ are essentially independent.

As the theorem below shows, the leading term in an expansion of bias is of size $(nl)^{-1}\lambda^2$, and equals $-\beta_1 + \beta_2 + o(\lambda^2/nl)$ where

$$\beta_1 \equiv 2(nl)^{-1}\sum_{i=1}^{\infty} i\gamma\left(\frac{i}{\lambda}\right) = \frac{\lambda^2}{nl}c_1 + o\left(\frac{\lambda^2}{nl}\right),$$

$$c_1 \equiv 2\int_0^{\infty} t\gamma(t)\,\mathrm{d}t,$$

$$\beta_2 \equiv 2\mathrm{E}\{p(1, 3)(\overline{X}_1 - \mu)(\overline{X}_3 - \mu)\}.$$

We shall also show that β_2 is typically of the same order as β_1.

Theorem 1. *Assume Conditions* (C_1) *on the process Y, and Conditions* (C_2) *on the matching rule, with* $\alpha > \max(8, 4/\epsilon)$, $\lambda = o(l)$ *and* $l = o(n)$. *Then*

$$\mathrm{E}(\tilde{\sigma}^2) = \sigma^2 - \beta_1 + \beta_2 + O(\lambda n^{-2} l) + o\left(\frac{\lambda^2}{nl}\right), \tag{5.2}$$

$$\mathrm{var}\,(\tilde{\sigma}^2) = 2n^{-1} l\sigma^4 + o\{\lambda^2 n^{-3} l + \lambda^4 (nl)^{-2}\}. \tag{5.3}$$

Since either β_1^2 or $n^{-1} l\sigma^4$ dominates each remainder term then it is always true that

$$\mathrm{E}\{(\tilde{\sigma}^2 - \sigma^2)^2\} \sim (\beta_1 - \beta_2)^2 + 2n^{-1} l\sigma^4.$$

If in addition $l = o\{(n\lambda)^{1/2}\}$ and $n\lambda^2 = O(l^3)$ then

$$\mathrm{E}(\tilde{\sigma}^2) - \sigma^2 \sim -\beta_1 + \beta_2, \qquad \mathrm{var}\,(\tilde{\sigma}^2) \sim 2n^{-1} l\sigma^4. \tag{5.4}$$

The last result represents an analogue of (3.2) and (3.3).

In order to compute the exact order of β_2 and its leading term, we need stronger assumptions on the matching rules. A class of different rules is covered by the following theorem.

Theorem 2. *Assume in addition to the conditions of the previous theorem that for some* ϵ_1, $\epsilon_2 > 0$, *and all* $1 < q < \infty$,

$$\|bp(1, 3) - I(|X_{1l} - X_{31}| \leqslant b^{-\epsilon_1})\{P(|X_{1l} - X_{31}| \leqslant b^{-\epsilon_1} | X_{1l})\}^{-1}\|_q = O(b^{-\epsilon_2}).$$

Suppose too that $\lambda \to \infty$, Y *is an almost surely continuous process and that* $Y(t)$ *has a continuous density with respect to Lebesgue measure. Then*

$$\beta_2 = \frac{\gamma_2 \lambda^2}{nl} + o\left(\frac{\lambda^2}{nl}\right), \tag{5.5}$$

where

$$\gamma_2 = \sum_{i=-l+1}^{0} \sum_{k=0}^{l-1} \mathrm{E}\left\{\mathrm{E}\left[\left\{Y\left(\frac{i}{\lambda}\right) - \mu\right\}\middle| Y(0)\right]\mathrm{E}\left[\left\{Y\left(\frac{k}{\lambda}\right) - \mu\right\}\middle| Y(0)\right]\right\}\lambda^{-2}.$$

If in addition the process Y *is Gaussian, with* $\gamma(t) - \gamma(0) = O(|t|^{\epsilon})$ *for some* $\epsilon > 0$, *then*

$$\gamma_2 \sim c_2 = 2\gamma(0)^{-1}\left(\int_0^\infty \gamma(t)\,\mathrm{d}t\right)^2. \tag{5.6}$$

Remark 1. Observe that the first-order contributions to squared bias and variance are of sizes $\lambda^4(nl)^{-2}$ and $\lambda^2 n^{-3} l$, respectively. Therefore the optimal block length is of size $(n\lambda^2)^{1/3}$. Result (5.4) holds for such values of l.

Remark 2. As in Section 3, $\gamma_2 \sim c_1$ if Y is a Markov process.

Remark 3. For a general stationary distribution π_i,

$$\mathrm{var}\,(S_1) \sim 2\sum \pi_i^2 \sigma^4,$$

so for (5.3) it seems necessary to have $\sum \pi_i^2 \sim b^{-1}$. This implies, via the Cauchy–Schwarz inequality, that the stationary distribution is approximately uniform.

Remark 4. Without Condition (C_{1A}) we should replace the indices 1 and 3 in β_2 by two indices j_1, j_2 with $|j_1 - j_2|$ tending to infinity. Since then $\mathscr{B}_{j_1}$ and $\mathscr{B}_{j_2}$ become independent, we obtain (3.7) from (3.4).

The effect of block matching on the bias may be studied most easily for Gaussian processes. Figure 3 depicts the asymptotic value of the ratio of the biases for matched and non-matched blocks, $1 - c_2/c_1$, for the covariance functions $\gamma(t) = \exp(-c|t|^\alpha)$ (In this example we do not adhere to Condition (C_{1A})). Here $0 < \alpha \leqslant 2$, and, the larger α, the smoother the process is. This example shows that the reduction in bias can be substantial.

To appreciate that block matching in terms of nearness of block ends is counter-productive for a time series with a considerable amount of repulsion, note that, because c_2 is always positive, block matching by nearness of block ends exacerbates the bias problem when $c_1 < 0$. To be specific, consider the case

$$\gamma(t) = \begin{cases} (1 - |t|)\cos(\omega t) & \text{if } |t| \leqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

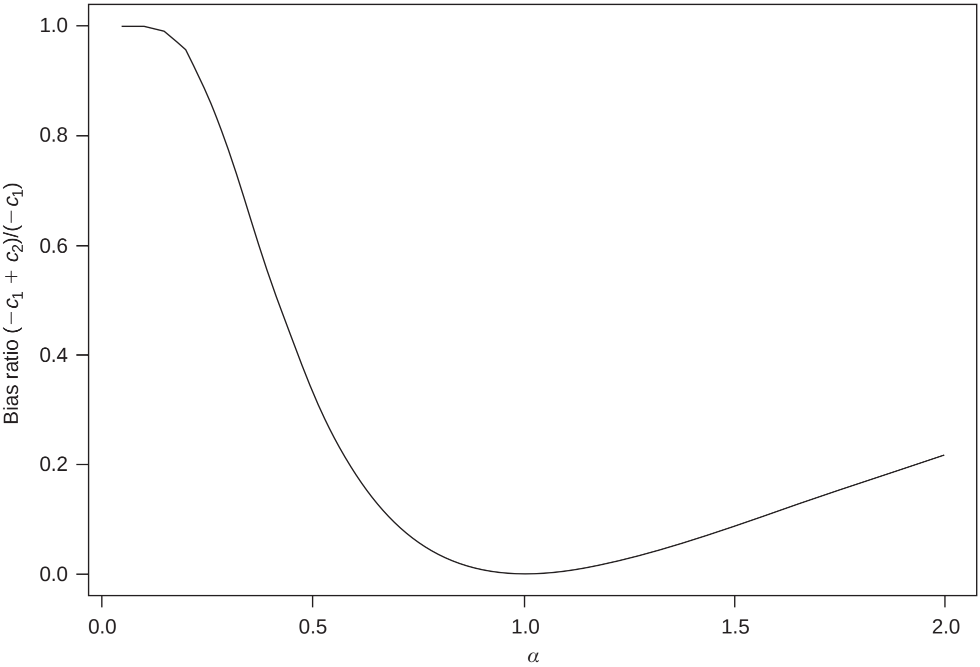

Fig. 3. Ratio of the bias of the matched to the bias of the non-matched block bootstrap. The figure depicts values of the ratio $1 - c_2/c_1$ for Gaussian processes with $\gamma(t) = \exp(-c|t|^\alpha)$, where $0 < \alpha \leqslant 2$; the ratio is independent of c.

where ω is a parameter of the process. The value of c_1 for this covariance function is $4\omega^{-3}\sin\omega - 2\omega^{-2}(1+\cos\omega)$, which is negative for many choices of ω. For $\omega \approx \pi$ the bias of the matched block estimator is substantially larger than the bias of the non-matched block bootstrap (Figure 4). Similar behaviour is observed with other covariance functions that have negative parts.

Appendix: Proofs of theorems

A.1. Preliminaries

Assume that $\mathrm{E}(Y) = 0$, and define

$$T_3 = \sum_{j=1}^{b} \overline{X}_j^2, \qquad T_4 = \sum_{j_1=1}^{b}\sum_{j_2=1}^{b} p(j_1, j_2)\overline{X}_{j_1}\overline{X}_{j_2}, \qquad T_5 = \sum_{j_1=1}^{b}\sum_{j_2=1}^{b} p(j_1, j_2)\overline{X}_{j_2}.$$

Since $\sum_{j_2} p(j_1, j_2) = 1$, then $T_2 = T_4 - \overline{X}T_5$. Therefore,

$$\tilde{\sigma}^2 = b^{-2}(T_3 + 2T_4) - b^{-1}\overline{X}^2 - 2b^{-2}\overline{X}T_5.$$

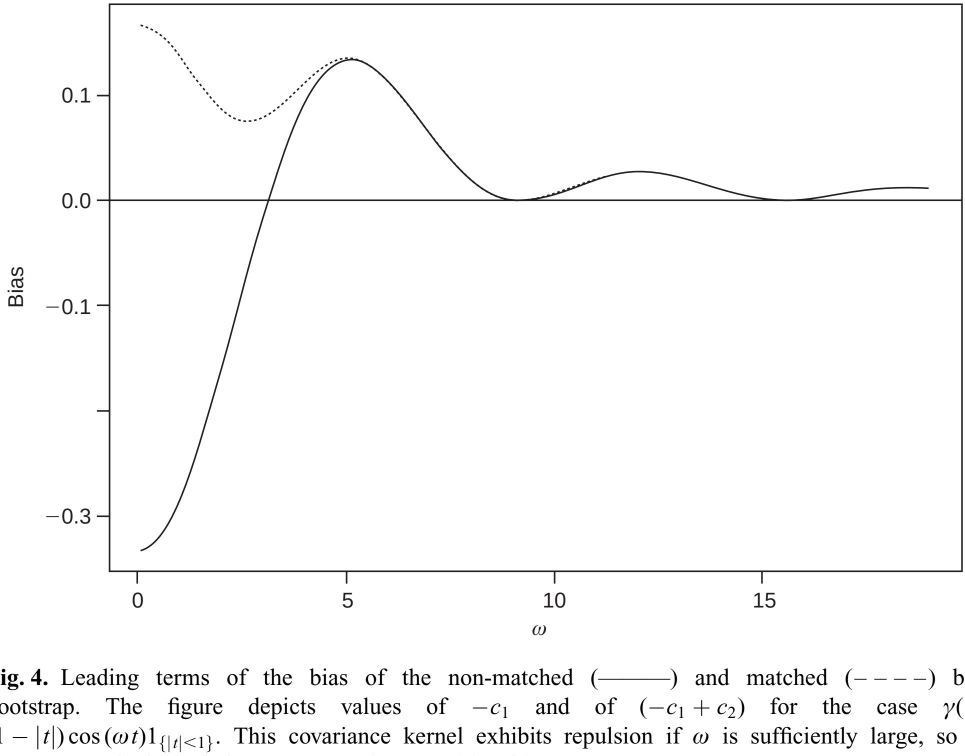

Fig. 4. Leading terms of the bias of the non-matched (———) and matched (– – – –) block bootstrap. The figure depicts values of $-c_1$ and of $(-c_1 + c_2)$ for the case $\gamma(t) = (1 - |t|)\cos(\omega t)1_{\{|t|<1\}}$. This covariance kernel exhibits repulsion if ω is sufficiently large, so that the non-matched block bootstrap can have positive bias.

It is straightforward to prove that

$$\mathrm{E}(\overline{X}^4) = O(b^{-2}l^{-2}\lambda^2),$$

and we shall show in Section A.2 that, for some $\zeta > 0$,

$$\mathrm{E}(\overline{X}^2 T_5^2) = O\left\{ b^{1-\zeta}\left(\frac{\lambda}{l}\right)^2 + b^{2-\zeta}\left(\frac{\lambda}{l}\right)^4 \right\}. \tag{A.1}$$

With similar but simpler arguments one can show that

$$\mathrm{E}(\overline{X}T_5) = O\left\{ \lambda l^{-1} + b^{1-\zeta}\left(\frac{\lambda}{l}\right)^2 \right\}.$$

In Section A.3 we show that

$$\mathrm{E}(T_4) = b^2\mathrm{E}\{p(1, 3)\overline{X}_1\overline{X}_3\} + O\left(\frac{\lambda}{l}\right) + O\left\{ b^{1-\zeta}\left(\frac{\lambda}{l}\right)^2 \right\}. \tag{A.2}$$

Using an argument similar to that in Section A.2 it may be proved that

$$\mathrm{var}\,(T_4) = o\left\{ b^4\left(\frac{\lambda}{n}\right)^2 (b^{-1} + l^{-2}\lambda^2) \right\}.$$

From the bootstrap with independent blocks we know that

$$\mathrm{E}(T_3) = b\sigma^2(l) = b^2\sigma^2(n) - b^2\beta_1 + O\left\{ \left(\frac{\lambda}{l}\right)^2 \right\},$$

$$\mathrm{var}\,(T_3) \sim 2b^3\sigma^4(n) = O\left\{ b\left(\frac{\lambda}{l}\right)^2 \right\}.$$

These results, together with the Cauchy–Schwarz inequality and the fact that $b = n/l$, imply (5.2) and (5.3).

A.2. Proof of (A.1)

Observe that

$$\mathrm{E}(\overline{X}^2 T_5^2) = b^{-2}\sum{}^{(6)}\mathrm{E}\{p(j_1, j_2)p(j_3, j_4)\overline{X}_{j_2}\overline{X}_{j_4}\overline{X}_{j_5}\overline{X}_{j_6}\} = b^{-2}l^{-4}S, \tag{A.3}$$

where

$$S \equiv \sum{}^{(10)}\mathrm{E}\{p(j_1, j_2)p(j_3, j_4)X_{j_2k_1}X_{j_4k_2}X_{j_5k_3}X_{j_6k_4}\},$$

the sixfold sum $\sum^{(6)}$ is over vectors $(j_1, \ldots, j_6) \in \{1, \ldots, b\}^6$, and the tenfold sum $\sum^{(10)}$ is over those vectors and also over $(k_1, \ldots, k_4) \in \{1, \ldots, l\}^4$.

We bound S by considering a number of different configurations of the vectors $(j_1, \ldots, j_6)$ and $(k_1, \ldots, k_4)$. We call k_i a *boundary index* if $k_i \leqslant r + t_0\lambda$ or $k_i \geqslant l - r - t_0\lambda + 1$, and an interior index otherwise. The cases identified below cover all distinct configurations up to isomorphisms. Since there is only a bounded number of the latter then we do not treat them here.

Case I. *$k_1, \ldots, k_4$ are all interior indices.* Here the term

$$\mathrm{E}\{p(j_1, j_2)p(j_3, j_4)X_{j_2k_1}X_{j_4k_2}X_{j_5k_3}X_{j_6k_4}\} \tag{A.4}$$

factorizes into the product of $\mathrm{E}\{p(j_1, j_2)p(j_3, j_4)\}$ and $\mathrm{E}(X_{j_2k_1}X_{j_4k_2}X_{j_5k_3}X_{j_6k_4})$. The second factor equals zero unless one of the following subcases holds, or one of the bounded number of possibilities isomorphic to these obtains.

Subcase (a). $j_2 = j_4 = j_5 = j_6$.

Subcase (b). $j_2 = j_5 \neq j_4 = j_6$.

Subcase (c). $j_2 = j_4 \neq j_5 = j_6$.

In Subcase (a) the sum over $k_1 \ldots, k_4$ contributes a term of order $l^2\lambda^2$, and the sum over j_1, j_2 and j_3 contributes another $O(b^2)$. (Note that the sum of $p(j_1, j_2)$ over its second index is identically 1.) Since the sums are in multiple, then these two contributions should be multiplied together, and so the contribution to S obtained by summing the term in (A.4) over indices corresponding to Subcase (a) is $b^2l^2\lambda^2$. The argument in Subcase (b) is similar, with identical orders of magnitude arising from summation over $k_1, \ldots, k_4$ and over $j_1, \ldots, j_4$. Therefore, the contribution to S that arises in Subcase (b) is again $O(b^2l^2\lambda^2)$.

The contribution to S from Subcase (c) is

$$\sum_{j_1}\sum_{j_2}\sum_{j_3}\sum_{j_5}\mathrm{E}\{p(j_1, j_2)p(j_3, j_2)\}O(l^2\lambda^2). \tag{A.5}$$

In bounding the expectation we may suppose that $j_2 \neq j_3 + 1$, since the contrary case may be treated more simply. (There, the number of sums in (A.5) is effectively only three, not four.) Under this assumption we may define $p'(j_1, j_2, j_3 + 1, j_3 + 1)$ as in Condition $(\mathrm{C}_{2\mathrm{D}})$. Then,

$$\begin{aligned}\mathrm{E}\{p(j_1, j_2)p(j_3, j_2)\} \leqslant {}& \mathrm{E}\{|p(j_1, j_2) - p'(j_1, j_2, j_3 + 1, j_3 + 1)|p(j_3, j_2)\}\\ &+ \mathrm{E}\{p'(j_1, j_2, j_3 + 1, j_3 + 1)p(j_3, j_2)\}.\end{aligned}$$

By the symmetry in Condition $(\mathrm{C}_{2\mathrm{A}})$, $\mathrm{E}\{p(j_3, j_2)\} \leqslant 1/(b-1)$. Using Condition $(\mathrm{C}_{2\mathrm{D}})$ and choosing any $q > 1$, the first term on the right is seen to be bounded by

$$\|p(j_1, j_2) - p'(j_1, j_2, j_3, j_3)\|_q\mathrm{E}\{p(j_3, j_2)\}^{1-1/q} = O(b^{-1-\epsilon}b^{-1+1/q}) = O(b^{-1-\epsilon}).$$

Moreover, by Conditions $(\mathrm{C}_{2\mathrm{D}})$ and $(\mathrm{C}_{2\mathrm{E}})$, the second term on the right is bounded by

$$\mathrm{E}[p'(j_1, j_2, j_3 + 1, j_3 + 1)\mathrm{E}\{p(j_3, j_2)|V_{j_3}; \mathscr{B}_j, j \neq j_3, j_3 + 1\}]$$

$$\begin{aligned}&\leqslant Cb^{-\epsilon}\mathrm{E}\{p'(j_1, j_2, j_3+1, j_3+1)\}\\ &\leqslant Cb^{-\epsilon}\mathrm{E}\{|p(j_1, j_2) - p'(j_1, j_2, j_3+1, j_3+1)| + p(j_1, j_2)\}\\ &\leqslant C'b^{-1-\epsilon}.\end{aligned}$$

Therefore, the expectation in (A.5) equals $O(b^{-1-\epsilon})$, and so the quantity in (A.5) equals $O(b^{3-\epsilon}l^2\lambda^2)$.

Combining the results from Subcases (a)–(c) we see that the contribution to S that arises from Case I equals $O(b^{3-\zeta}l^2\lambda^2)$, for some $\zeta > 0$.

Case II. *$k_1, \ldots, k_4$ are all boundary indices.* Defining

$$\pi = \pi(j_1, \ldots, j_4) = p(j_1, j_2)p(j_3, j_4),$$

the term in (A.4) becomes $\mathrm{E}(\pi X_{j_2k_1}X_{j_4k_2}X_{j_5k_3}X_{j_6k_4})$. We consider separately the following subcases:

Subcase (a). j_5 or j_6 belongs to $\{j_1-1, j_1, j_2, j_3-1, j_3, j_4\}$.

Subcase (b). This covers all other situations.

In Subcase (a) we bound the term by

$$\{\mathrm{E}(\pi^{\alpha/(\alpha-4)})\}^{(\alpha-4)/\alpha}(\mathrm{E}|X_{j_2k_1}X_{j_4k_2}X_{j_5k_3}X_{j_6k_4}|^{\alpha/4})^{4/\alpha}.$$

In Subcase (a) the number of values (j_5, j_6) is $O(b)$ uniformly in $(j_1, \ldots, j_4)$, so summing over j_5 and j_6 gives a contribution $O(b)$. By Hölder's inequality

$$\begin{aligned}\sum_{j_1\cdots j_4}\|\pi\|_{\alpha/(\alpha-4)} &\leqslant \left(\sum_{j_1\cdots j_4}\mathrm{E}(\pi^{\alpha/(\alpha-4)})\right)^{(\alpha-4)/\alpha}\left(\sum_{j_1\cdots j_4}1\right)^{4/\alpha}\\ &\leqslant \left(\sum_{j_1\cdots j_4}\mathrm{E}(\pi)\right)^{(\alpha-4)/\alpha} b^{16/\alpha}\\ &= b^{2+8/\alpha}.\end{aligned}$$

Combining these results we see that the total contribution to S in Subcase (a) equals $O(b^{3+8/\alpha}\lambda^4)$. Taking $\alpha > 8$ thus ensures that this contribution does not exceed

$$O(b^{4-\zeta}\lambda^4) \tag{A.6}$$

for some $\zeta > 0$.

Next we treat Subcase (b). Let k_3 and k_4 be distant $O(\lambda)$ from $(j_5-1)l+1$ and $(j_6-1)l+1$, respectively. Define

$$\pi' = \pi'(j_1, \ldots, j_6) = p'(j_1, j_2, j_5, j_6)p'(j_3, j_4, j_5, j_6).$$

Then we have, in view of the independence of $X_{j_5k_3}X_{j_6k_4}$ and $\pi' X_{j_2k_1}X_{j_4k_2}$,

$$|\mathrm{E}(\pi X_{j_2k_1}X_{j_4k_2}X_{j_5k_3}X_{j_6k_4}) - \mathrm{E}(\pi X_{j_2k_1}X_{j_4k_2})\mathrm{E}(X_{j_5k_3}X_{j_6k_4})|$$
$$\leqslant |\mathrm{E}\{(\pi - \pi')X_{j_2k_1}X_{j_4k_2}X_{j_5k_3}X_{j_6k_4}\}| + \mathrm{E}(X_1^2)|\mathrm{E}\{(\pi - \pi')X_{j_2k_1}X_{j_4k_2}\}|. \tag{A.7}$$

By Hölder's inequality, the first term on the right-hand side of (A.7) is bounded by $\|\pi - \pi'\|_{\alpha/(\alpha-4)}\|X_1\|_\alpha^4$. Since

$$\pi - \pi' = p(j_1, j_2)\{p(j_3, j_4) - p'(j_3, j_4, j_5, j_6)\} + p(j_3, j_4)\{p(j_1, j_2) - p'(j_1, j_2, j_5, j_6)\}$$
$$- \{p(j_1, j_2) - p'(j_1, j_2, j_5, j_6)\}\{p(j_3, j_4) - p'(j_3, j_4, j_5, j_6)\},$$

then the triangle inequality, Hölder's inequality and Condition ($\mathrm{C}_{2\mathrm{D}}$) imply that, for any $\xi > 1$,

$$\|\pi - \pi'\|_{\alpha/(\alpha-4)} = O(b^{-1-\epsilon})\{\|p(j_1, j_2)\|_{\xi\alpha/(\alpha-4)} + \|p(j_3, j_4)\|_{\xi\alpha/(\alpha-4)}\} + O(b^{-2(1+\epsilon)}).$$

Similarly, the second term on the right-hand side of (A.7) is bounded by

$$\|\pi - \pi'\|_{\alpha/(\alpha-2)}\|X_1\|_\alpha^2\|X_1\|_2^2 = O(b^{-1-\epsilon})\{\|p(j_1, j_2)\|_{\xi\alpha/(\alpha-2)} + \|p(j_3, j_4)\|_{\xi\alpha/(\alpha-2)}\}$$
$$+ O(b^{-2(1+\epsilon)}).$$

If k_3 is within distance $O(\lambda)$ of $j_5 l$, then we argue as above but with the definition of π' altered to $p'(j_1, j_2, j_5 + 1, j_6)p'(j_3, j_4, j_5 + 1, j_6)$. Thus, the bounds just derived hold for any of the terms arising in Subcase (b) of Case II. Moreover, $\mathrm{E}(X_{j_5k_3}X_{j_6k_4}) = 0$ unless $|j_5 - j_6| \leqslant 1$. This, together with (A.7) and the bounds above, produce the following bound for the contribution to S from Case II, Subcase (b):

$$O(b\lambda^4)\sum_{j_1,\ldots,j_4}\|\pi(j_1, j_2, j_3, j_4)\|_{\alpha/(\alpha-2)} + O(b^{4-(1+\epsilon)}\lambda^4)\sum_{j_1,j_2}\|p(j_1, j_2)\|_{\xi\alpha/(\alpha-4)} + O(b^{6-2(1+\epsilon)}\lambda^4).$$

Arguing as in the derivation of the bound at (A.6) we see that this equals

$$O(b^{3+4/\alpha}\lambda^4 + b^{5+4\xi/\alpha-\epsilon-1/\xi}\lambda^4 + b^{4-2\epsilon}\lambda^4),$$

which, since we made the assumption that $\alpha > 4/\epsilon$, may be rendered of the order at (A.6) by choosing $\xi > 1$ sufficiently close to 1.

Adding the bounds from Subcases (a) and (b) we see that the total contribution to S from terms considered under Case II is of the order in (A.6).

Case III. *Three k_i values are boundary indices and the other is interior.* Here the contribution to S is identically zero.

The methods used to derive the bounds in Cases IV–VII below are somewhat different and are given only in barest outline here. Although the bounds are identical, none of the cases is isomorphic to another.

Case IV. *k_1, k_2 are boundary indices and k_3, k_4 are interior.* The contribution is identically zero unless $j_5 = j_6$, and there the contributions from the sums over

$$(k_3, k_4), (k_1, k_2), j_5 \text{ and } (j_1, j_2, j_3, j_4)$$

are $O(l\lambda)$, $O(\lambda^2)$, $O(b)$ and $O(b^{2(\alpha-2)/\alpha}b^{4(2/\alpha)})$, respectively. Multiplying them together we see that the total contribution to S is $o(b^{3+4/\alpha}l\lambda^3)$. Because $\alpha > 8$ and the geometric mean is bounded by the arithmetic mean, we have that $b^{3+4/\alpha}l\lambda^3 = O(b^{3-\zeta}l^2\lambda^2 + b^{4-\zeta}\lambda^4)$ for some $\zeta > 0$.

Case V. *k_3, k_4 are boundary indices and k_1, k_2 are interior.* The contribution to S is identically zero unless $j_2 = j_4$, and the contribution from the latter source is $O(b^{4-\zeta}l\lambda^3)$ for some $\zeta > 0$, using an argument similar to that employed to treat Subcase (c) of Case I.

Case VI. *k_1, k_3 are boundary indices and k_2, k_4 are interior indices.* The contribution to S is identically zero unless $j_4 = j_6$, and the contribution from the latter source is $O(b^{3+4/\alpha}l\lambda^3)$.

Case VII. *k_4 is a boundary index and the others are all interior.* The contribution to S is identically zero unless $j_2 = j_4 = j_5$, and the contribution from the latter source is $O(b^{3+1/\alpha}l\lambda^3)$.

Case VIII. *k_2 is a boundary index and the others are all interior.* The contribution to S is identically zero unless $j_2 = j_5 = j_6$, and the contribution from the latter source is $O(b^{2+2/\alpha}l\lambda^3)$.

Now we add the bounds derived in each of the eight cases. Because $\alpha > 8$ and the geometric mean is bounded by the arithmetic mean, $b^{3+4/\alpha}l\lambda^3 = O(b^{3-\zeta}l^2\lambda^2 + b^{4-\zeta}\lambda^4)$ for some $\zeta > 0$. Hence we obtain that, for some $\zeta > 0$,

$$S = O(b^{3-\zeta}l^2\lambda^2 + b^{4-\zeta}\lambda^4).$$

Result (A.1) now follows from (A.3).

A.3. Calculation of E(T_4)

Note that, by symmetry, $\mathrm{E}\{p(j_1, j_2)\overline{X}_{j_1}\overline{X}_{j_2}\}$ is the same for any $j_1 < b$, $j_2 < b$, $|j_2 - j_1| > 1$. By an argument similar to that in Section A.2 we may show that, for any j_1, j_2,

$$|\mathrm{E}\{p(j_1, j_2)\overline{X}_{j_1}\overline{X}_{j_2}\}| = O\left(\frac{\lambda}{l}\right)\mathrm{E}\{p(j_1, j_1)\}\delta_{j_1,j_2} + O\left\{\left(\frac{\lambda}{l}\right)^2\right\}\|p(j_1, j_2)\|_{\alpha/(\alpha-2)}.$$

Since $p(j_1, j_2) \leqslant 1$, then

$$\|p(j_1, j_2)\|_{\alpha/(\alpha-2)} \leqslant [\mathrm{E}\{p(j_1, j_2)\}]^{(\alpha-2)/\alpha}.$$

Using the fact that $\mathrm{E}\{p(j_1, j_2)\} \leqslant 1/(b-1)$ if $j_2 \neq j_1 + 1$, and Condition ($\mathrm{C}_{2\mathrm{C}}$) if $j_2 = j_1 + 1$, we obtain, for $\alpha > 8$,

$$\mathrm{E}(T_4) - b^2\mathrm{E}\{p(1, 3)\overline{X}_1\overline{X}_3\} = O\left(\frac{\lambda}{l}\right) + O\left\{\left(\frac{\lambda}{l}\right)^2 b^{1-3\epsilon/4}\right\},$$

which is (A.2).

A.4. Proof of (5.5) and (5.6)

Let Y, Y_1, Y_2 be independent processes with identical laws, and put

$$Z_j = \sum_{i=0}^{l} Y_j\left(\frac{i}{\lambda}\right).$$

It is easy to see that, under the conditions of Theorem 2,

$$\beta_2 = b^{-1}l^{-2}\mathrm{E}[\mathrm{E}\{Z_1 Z_2 | Y_1(0) = Y_2(0)\}] + o\left(\frac{\lambda^2}{bl^2}\right),$$

which is (5.5). Finally, for a Gaussian process, $\mathrm{E}\{Y(i/\lambda)|Y(0)\} = \gamma(i/\lambda)/\gamma(0)Y(0)$, which gives (5.6).

Acknowledgements

The helpful comments of two referees are gratefully acknowledged.

References

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

Bühlmann, P. (1995) The blockwise bootstrap for general empirical processes of stationary sequences. *Stoch. Processes Applic.*, **58**, 247–265.

Bühlmann, P. and Künsch, H.R. (1994) Block length selection in the bootstrap for time series. Research Report 72, Seminar für Statistik, Eidgenössische Technische Hochschule Zurich.

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

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.

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

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

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

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

Hall, P. (1985) Resampling a coverage pattern. *Stoch. Processes Applic.*, **20**, 231–246.

Hall, P. and Jing, B. (1996) On sample reuse methods for dependent data. *J. Roy. Statist. Soc., Ser. B*, **58**, 727–737.

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

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

Lahiri, S.N. (1991) Second-order optimality of stationary bootstrap. *Statist. Probab. Lett.*, **11**, 335–341.

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.

Liu, R. and Singh, K. (1992) Moving blocks jackknife and bootstrap capture weak dependence. In R. LePage and L. Billard (eds), *Exploring the Limits of the Bootstrap*, pp. 225–248. New York: Wiley.

Naik-Nimbalkar, U.V. and Rajarshi, M.B. (1994) Validity of blockwise bootstrap for empirical processes with stationary observations. *Ann. Statist.*, **22**, 980–994.

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 spectral density estimation. *Ann. Statist.*, **20**, 1985–2007.

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

Politis, D.N. and Romano, J.P. (1995) Bias-corrected nonparametric spectral estimation. *J. Time Series Anal.*, **16**, 67–103.

Radulovic, D. (1995) The bootstrap of empirical processes for α-mixing sequences. Preprint, University of Connecticut.

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

Radulovic, D. (1996b) The bootstrap for empirical processes based on stationary observations. *Stoch. Processes Applic.*, **65**, 259–279.

Shao, Q.M. and Yu, H. (1993) Bootstrapping the sample means for stationary mixing sequences. *Stoch. Processes Applic.*, **48**, 175–190.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Wood, A.T.A. and Chan, G. (1994) Simulation of stationary Gaussian processes. *J. Comput. Graph. Statist.*, **31**, 409–432.

Received January 1996 and revised November 1996