Sieve bootstrap with variable length Markov chains for stationary categorical time series

Peter Bühlmann
Seminar für Statistik, ETH Zürich, Switzerland

2nd revision
May 2001

Abstract

We study a bootstrap for stationary categorical time series based on the method of sieves. The data generating process is approximated by so-called variable length Markov chains [VLMC], a flexible class of Markov models which allows for parsimonious structure. Then, the resampling is given by simulating from the fitted model.

We show that for a whole class of stationary categorical time series, which is more general than VLMC's, the VLMC-sieve has faster rate of convergence for variance estimation than the more general block bootstrap. Our results are illustrated from a theoretical and empirical perspective. For the latter, we also present a real data application about (in-) homogeneity classification of a DNA strand.

Finally, the VLMC-sieve scheme enjoys an implementational advantage of using the plug-in rule for bootstrapping a statistical procedure, which is generally not the case for the block method.

Key Words. Binary sequence, block bootstrap, confidence interval, DNA sequence, double bootstrap, variance estimation.

Heading: VLMC-sieve bootstrap

1 Introduction

Estimating standard errors or the distribution of a statistical procedure in time series is a difficult task due to dependence among observations. Already the asymptotic variance of a mean is generally infinite-dimensional and cannot be estimated with convergence rate $1/\sqrt{\text{sample size}}$. Bootstrap techniques ease some of the difficulties and are a very important tool.

The block bootstrap introduced by Künsch (1989) is a procedure designed for very general stationary processes. For the special case of categorical time series, we study here a bootstrap based on the method of sieves. That is, the data generating mechanism is approximated by a family of parametric models such that in the limit as sample size tends to infinity, the true underlying process is contained [in some sense] in the sequence of approximating models. For finite sample size n, the approximating models are typically misspecified. The resampling is then based on a fitted model [an element of the sieve], selected from the data. We use here the sieve of so-called variable length Markov chains [VLMC], a flexible class of potentially high order Markov models allowing for parsimonious structure. For more details about VLMC's, see Bühlmann and Wyner (1999), where also bootstrapping in a VLMC has been considered. Here, the main new issues are:

(i) the VLMC-based bootstrap can be viewed as a sieve bootstrap for *general* categorical time series, which are not necessarily a VLMC;

(ii) the VLMC-sieve bootstrap is often *more accurate* than the block bootstrap. We prove an asymptotic result for variance estimation if the underlying process has exponentially fast decaying strength of dependence as separation lags grow.

There seems to be no other result in this area. For a fixed finite state Markov chain model, accuracy of a Markov model-based bootstrap has been theoretically analyzed in Datta and McCormick (1995) which is an answer for (ii) without allowing for general categorical time series. Other references for the bootstrap in a finite state Markovian model include as one of the earliest papers Kulperger and Prakasa Rao (1989). Both of these and other related results about [full] Markov chains are restricted to a fixed model and are thus sensitive to model-misspecification. And unlike a result about sieve approximation, they don't indicate accuracy outside the model, describing why and when they are expected to work well and should be preferred over the block bootstrap. There is thus a substantial theoretical and also empirical gap about the accuracy of sieve approximation with Markov chains. We give here an answer to this gap for categorical time series. Both issues (i) and (ii) from above are illustrated from an asymptotic theoretical as well as a finite sample practical view. For the latter, we give some results from a simulation study as well as an application for the search of '$C\mathbf{p}G$-islands' which are indicators for starting regions of genes [inhomogeneity regions] in a long DNA sequence, cf. Bird (1987).

An important implementational advantage of the VLMC-sieve bootstrap is given by its definition with the practically convenient plug-in rule: the same statistical procedure or program $h_n(.)$ for the original sample $X_1, \ldots, X_n$ is applied to the bootstrap resample $X_1^*, \ldots, X_n^*$. This has its price: the construction of the bootstrap resample requires a sophisticated fitting algorithm. The difficulty should become less pronounced when using our publicly available software which is also included in the statistical computing language

R, see Appendix A. On the other hand, the resampling for the block bootstrap is very easy to implement but then, the method does not use the plug-in rule so that the computation of the block bootstrapped estimator may have to be redesigned.

Example 1.1. Consider binary data $X_1, \ldots, X_n$. Let

$$\hat{\theta}_n = h_n(X_1, \ldots, X_n) = \frac{\sum_{t=1}^{n-1} 1_{[(X_t, X_{t+1})=(0,1)]}}{\sum_{t=1}^{n} 1_{[X_t=0]}},$$

the relative frequency of symbol '1' given the previous symbol '0', which is a sensible estimate for $\mathbb{P}[X_t = 1|X_{t-1} = 0]$, provided that the data-generating process is stationary. Block bootstrapping this estimator involves the following steps:
(1) Vectorize the observations $Y_1 = (X_1, X_2), Y_2 = (X_2, X_3), \ldots, Y_{n-1} = (X_{n-1}, X_n)$.
(2) Run the block resampling on the vectorized observations. Build overlapping blocks of length ℓ of consecutive vectorized observations. Then resample blocks independently with replacement, yielding

$$Y^*_{S_1+1}, \ldots, Y^*_{S_1+\ell}, Y^*_{S_2+1}, \ldots, Y^*_{S_2+\ell}, \ldots, Y^*_{S_k+1}, \ldots, Y^*_{S_k+\ell}.$$

This is re-denoted in short, using chronological order, by $Y_1^*, \ldots, Y_{n-1}^*$. Here, the block starting-points S_i are i.i.d. Uniform on $\{0, \ldots, n-\ell\}$ and for simplicity we assumed that $n - 1 = k\ell$.
(3) Approximate the original estimator by

$$\tilde{\theta}_n = \tilde{h}_n(Y_1, \ldots, Y_{n-1}) = \frac{\sum_{t=1}^{n-1} 1_{[Y_t=(0,1)]}}{\sum_{t=1}^{n-1} 1_{[(Y_t)_1=0]}} = \hat{\theta}_n + \text{ edge effects.}$$

The block bootstrap sample is now applied to $\tilde{\theta}_n$, yielding

$$\hat{\theta}_n^{*\text{block boot}} = \tilde{h}_n(Y_1^*, \ldots, Y_{n-1}^*).$$

This is the plug-in rule applied to $\tilde{h}_n(\cdot)$ with the vectorized block bootstrap sample. Would we simply run the block bootstrap on the non-vectorized observations and plugging them into $h_n(\cdot)$ [defining the original estimator $\hat{\theta}_n$], the numerator in $h_n(\cdot)$ would be evaluated at many two-dimensional points (X_t^*, X_{t+1}^*) which are not occurring in the original sample: this is due to the fact that joining resampled blocks creates very strange artificial pairs of observations. These artifacts are corrected when using vectorized observations and plugging into $\tilde{h}_n(\cdot)$. The procedure just described is in fact Künsch's (1989) original proposal, often bringing substantial improvements over the non-vectorized, 'naive' block bootstrap scheme which is a bad version of block bootstrap. But of course, the price to apply the block bootstrap correctly is the complicated implementational task in Step (3) above. While this is still an easy task in this Example 1.1, it could become much more awkward for more complicated estimators where the parameter of interest depends on the p-dimensional marginal distribution of the stochastic process; for Example 1.1, $p = 2$.

The new theoretical and empirical results in this paper and implementational issues are good reasons to consider the VLMC-sieve bootstrap as an interesting alternative resampling scheme for categorical time series.

2 The VLMC-sieve bootstrap for stationary categorical time series

Consider a sample realization $X_1, \ldots, X_n$ from a stationary categorical time series with values in a finite space $\mathcal{X}$. Given an estimator $\hat{\theta}_n = h_n(X_1, \ldots, X_n)$, we want to estimate its distribution with a bootstrap technique. We focus primarily on the class of statistics which are representable as a smooth functional of the empirical distribution,

$$\hat{\theta}_n = T(\hat{P}^{(m)}_{n,emp}),$$
$$\hat{P}^{(m)}_{n,emp}(x) = (n-m+1)^{-1} \sum_{t=1}^{n-m+1} 1_{[(X_t,\ldots,X_{t+m-1})=x]} \; (x \in \mathcal{X}^m, \; m \in \mathbb{N}) \tag{2.1}$$

with T a smooth functional, for simplicity here assumed to be real-valued.

The class of estimators in (2.1) is broad: it includes very many statistics for estimating the parameter

$$\theta = T(P^{(m)}), \; P^{(m)}(x) = \mathbb{P}[(X_1, \ldots, X_m) = x] \; (x \in \mathcal{X}^m),$$

depending on the m-dimensional marginal distribution of the process, where m is finite. An example is the smooth function of means,

$$\hat{\theta}_n = f\left((n-m+1)^{-1} \sum_{t=1}^{n-m+1} g(X_t, \ldots, X_{t+m-1})\right), \tag{2.2}$$

where $g : \mathcal{X}^m \to \mathbb{R}^q$ $(q \geq 1)$ is arbitrary and $f : \mathbb{R}^q \to \mathbb{R}$ continuously differentiable at $\mu_g = \mathbb{E}[g(X_t, \ldots, X_{t+m-1})]$. For example, estimators of transition probabilities in full Markov chains of order $m-1$ or other functions of frequencies of tuples up to size m, e.g. score and scan statistics in genetics [cf. Braun and Müller (1998) and the references in Section 4.3]. Other examples for (2.1) are maximum likelihood estimators in generalized linear models of autoregressive type with quite general link functions [cf. Fahrmeir and Tutz, 1994].

2.1 The sieve with variable length Markov chains

We propose a bootstrap which is based on an estimated sieve approximation of the true underlying process $(X_t)_{t\in\mathbb{Z}} \sim P$. The construction of the sieve is done with so-called variable length Markov chains. Denote by $x_i^j = x_j, x_{j-1}, \ldots, x_i$ $(i < j, \; i, j \in \mathbb{Z} \cup \{-\infty, \infty\})$ a vector whose components are written in reverse order, $wu = (w_{\mathrm{card}(w)}, \ldots, w_2, w_1, u_{\mathrm{card}(u)}, \ldots, u_2, u_1)$ is the concatenation of the vectors w and u; here and in the sequel, $\mathrm{card}(\cdot)$ denotes the cardinality of a vector or a set. Also, capital letters X are usually used for random variables and small letters x for deterministic values.

Definition 2.1 *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process with values $X_t \in \mathcal{X}$. Denote by $c : \mathcal{X}^\infty \to \cup_{j=0}^\infty \mathcal{X}^j \cup \mathcal{X}^\infty$ $(\mathcal{X}^0 = \emptyset)$ a (variable projection) function which maps*

$$c : x^0_{-\infty} \mapsto x^0_{-\ell+1}, \; \text{where } \ell \text{ is defined by}$$
$$\ell = \ell(x^0_{-\infty}) = \min\{k; \mathbb{P}[X_1 = x_1 | X^0_{-\infty} = x^0_{-\infty}] = \mathbb{P}[X_1 = x_1 | X^0_{-k+1} = x^0_{-k+1}] \text{ for all } x_1 \in \mathcal{X}\},$$

where $\ell \equiv 0$ corresponds to independence.

Then, $c(.)$ is called a context function and for any $t \in \mathbb{Z}$, $c(x_{-\infty}^{t-1})$ is called the context (the relevant past) at time t. Let $0 \le p \le \infty$ be the smallest integer such that

$$\mathit{card}(c(x_{-\infty}^0)) = \ell(x_{-\infty}^0) \le p \textit{ for all } x_{-\infty}^0 \in \mathcal{X}^\infty.$$

The number p is called the order of the context function $c(.)$, and if $p < \infty$, $(X_t)_{t\in\mathbb{Z}}$ is called a stationary variable length Markov chain [VLMC] of order p.

Due to stationarity of $(X_t)_{t\in\mathbb{Z}}$, transition probabilities are homogeneous in time and the restriction to indices $0, -1, \ldots$ in the definitions above is without loss of generality. We sometimes identify a VLMC $(X_t)_{t\in\mathbb{Z}}$ with its probability distribution P_c on $\mathcal{X}^\mathbb{Z}$, where the subscript indicates the functional dependence on the context function $c(.)$. Clearly, a VLMC of order p is a Markov chain of order p, with the additional structure of having a *memory of variable length ℓ*. If the context function $c(.)$ of order p is the full projection $x_{-\infty}^0 \mapsto x_{-p+1}^0$ for all $x_{-\infty}^0$, the VLMC is a full Markov chain of order p. A VLMC has an important representation as a graphical tree model, see Figure 2.1.

Definition 2.2 *Let $c(.)$ be a context function of a stationary VLMC. The ($\mathit{card}(\mathcal{X})$-ary) context tree τ is defined as*

$$\tau = \tau_c = \{w; w = c(x_{-\infty}^0),\ x_{-\infty}^0 \in \mathcal{X}^\infty\}.$$

The context function $c(.)$ can be reconstructed from τ_c. The context tree τ_c, which does not have to be complete with $\mathrm{card}(\mathcal{X})$ offsprings per internal node, is nothing else than the minimal state space of the VLMC P_c.

Example 2.1. $\mathcal{X} = \{0, 1\}$, order $p = 3$.
The function

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

can be represented by the tree τ_c on the left hand side in Figure 2.1. A 'growing to the left' sub-branch represents the symbol 0 and vice versa for the symbol 1. The state space is given by the terminal nodes $\{0, 100, 101, 11\}$ of the tree [read top down].

Example 2.2. $\mathcal{X} = \{0, 1, 2, 3\}$, order $p = 2$.
The function

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

can be represented by the tree τ_c on the right hand side in Figure 2.1. The minimal state space $\{0, 1, 2, 3, 33\}$ is again given by the tree [read top down]: note that the state 3 is an internal node. An alternative representation of the state 3 is given by the round-edged rectangle symbolizing the absent nodes 0,1 and 2 in depth 2, which can be thought as a completion of the tree with nodes lumped together to the state 3.

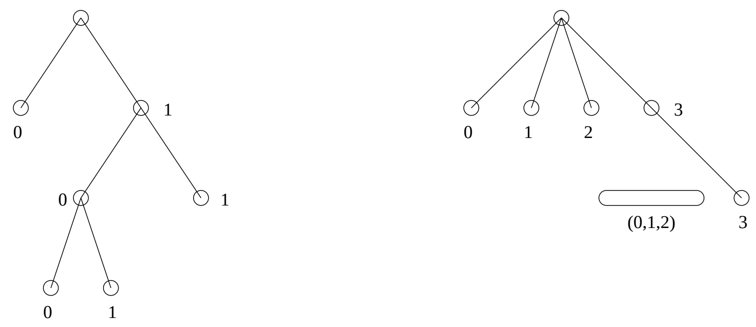

Figure 2.1: Tree representations of the variable length memories in Examples 2.1 and 2.2.

2.2 The context algorithm for fitting VLMC's

In the sequel, we often write for a probability measure P on $\mathcal{X}^{\mathbb{Z}}$, $P(x) = \mathbb{P}_P[X_1^q = x]$ $(x \in \mathcal{X}^q)$ [abbreviating $P^{(q)}(x)$] and $P(x|w) = P(xw)/P(w)$ $(x, w \in \cup_{j=1}^{\infty}\mathcal{X}^j)$. Fitting a VLMC involves a version of the tree structured context algorithm [Rissanen, 1983] for estimating the variable length memory, described by the context function $c(.)$, and for the set of transition probabilities. Estimation of τ_c is a highly complex model selection problem; due to the extremely large number of possible models, a natural tree hierarchy is employed. Once an estimate $\hat{\tau}$ is available, transition probabilities are determined by relative frequencies.

The tree structured context algorithm uses the notion of terminal node context trees

$$\tau^T = \tau_c^T = \{w; w \in \tau_c \text{ and } wu \notin \tau_c \text{ for all } u \in \mathcal{X}\}.$$

In Example 2.2, $\tau^T = \{0, 1, 2, 33\}$ is the set of terminal nodes in the tree in Figure 2.1, whereas $\tau = \tau^T \cup \{3\}$. In Example 2.1, τ^T and τ coincide. The information of τ^T is equivalent to the information in τ: the terminal node tree thus yields a more compact representation. Denote by

$$N(w) = \sum_{t=1}^{n-\mathrm{card}(w)+1} 1_{[X_t^{t+\mathrm{card}(w)-1}=w]}, \; w \in \cup_{m=1}^{\infty}\mathcal{X}^m, \tag{2.3}$$

the number of occurrences of the string w in the sequence X_1^n. Moreover, let

$$\hat{P}(w) = N(w)/n, \; \hat{P}(x|w) = \frac{N(xw)}{N(w)}, \; x, w \in \cup_{m=1}^{\infty}\mathcal{X}^m. \tag{2.4}$$

The algorithm below constructs the estimated context tree $\hat{\tau}$ as the biggest context tree (with respect to the order '$\preceq$' defined in Step 1 below) such that

$$\Delta_{wu} = \sum_{x\in\mathcal{X}} \hat{P}(x|wu)\log(\frac{\hat{P}(x|wu)}{\hat{P}(x|w)})N(wu) \geq K \text{ for all } wu \in \hat{\tau}^T \;\; (u \in \mathcal{X}) \tag{2.5}$$

with $K = K_n \sim C\log(n),\; C > 2\mathrm{card}(\mathcal{X}) + 4$ a cutoff to be chosen by the user.

Step 1 Given $\mathcal{X}$-valued data $X_1, \dots, X_n$, fit a maximal context tree, i.e., search for the context function $c_{max}(.)$ with terminal node context tree representation τ^T_{max}, where τ^T_{max} is the biggest tree such that every element (terminal node) in τ^T_{max} has been observed at least twice in the data. This can be formalized as follows:
τ^T_{max} is such that $w \in \tau^T_{max}$ implies $N(w) \geq 2$, and such that for every τ^T, where $w \in \tau^T$ implies $N(w) \geq 2$, it holds that $\tau^T \preceq \tau^T_{max}$.
Here, $\tau_1 \preceq \tau_2$ means: $w \in \tau_1 \Rightarrow wu \in \tau_2$ for some $u \in \cup_{m=0}^{\infty} \mathcal{X}^m$ $(\mathcal{X}^0 = \emptyset)$.
Set $\tau^T_{(0)} = \tau^T_{max}$.

Step 2 Examine every element (terminal node) of $\tau^T_{(0)}$ as follows (the order of examining is irrelevant). Let $c(.)$ be the corresponding context function of $\tau^T_{(0)}$ and let

$$wu = x^0_{-\ell+1} = c(x^0_{-\infty}),\ u = x_{-\ell+1},\ w = x^0_{-\ell+2},$$

where wu is an element (terminal node) of $\tau^T_{(0)}$, which we compare with its pruned version $w = x^0_{-\ell+2}$ (if $\ell = 1$, the pruned version is the empty branch $\emptyset$, i.e., the root node).
Prune $wu = x^0_{-\ell+1}$ to $w = x^0_{-\ell+2}$ if

$$\Delta_{wu} = \sum_{x \in \mathcal{X}} \hat{P}(x|wu) \log(\frac{\hat{P}(x|wu)}{\hat{P}(x|w)}) N(wu) < K,$$

with $K = K_n \sim C\log(n)$, $C > 2\text{card}(\mathcal{X}) + 4$ and $\hat{P}(.|.)$ as defined in (2.4). Decision about pruning for every terminal node in $\tau^T_{(0)}$ yields a (possibly) smaller tree $\tau_{(1)} \preceq \tau^T_{(0)}$. Construct the terminal node context tree $\tau^T_{(1)}$.

Step 3 Repeat Step 2 with $\tau_{(i)}, \tau^T_{(i)}$ instead of $\tau_{(i-1)}, \tau^T_{(i-1)}$ $(i = 1, 2, \dots)$ until no more pruning is possible. Denote this maximal pruned context tree (not necessarily of terminal node type) by $\hat{\tau} = \tau_{\hat{c}}$ and its corresponding context function by $\hat{c}(.)$.

Step 4 If interested in probability distributions, estimate the transition probabilities $P(x_1|c(x^0_{-\infty}))$ by $\hat{P}(x_1|\hat{c}(x^0_{-\infty}))$, where $\hat{P}(.|.)$ is defined as in (2.4).

More details and motivation can be found in Bühlmann and Wyner (1999).

The choice of the cutoff parameter $K = K_n$ in practice is often more intuitive on the scale of χ^2-quantiles. The decision statistic Δ_{wu} in Step 2 is a log-likelihood ratio statistic, having an asymptotic $\chi^2_{\text{card}(\mathcal{X})-1}/2$ distribution [under the null-hypothesis of equal models, i.e. pruning]. Therefore, we usually denote a cutoff K by

$$\alpha\% \text{ cutoff whose value is } \chi^2_{\text{card}(\mathcal{X})-1,\alpha}/2 = (\alpha\text{-quantile of } \chi^2_{\text{card}(\mathcal{X})-1})/2. \quad (2.6)$$

The choice in (2.6) does not reflect dependence on sample size n, but incorporates the size $\text{card}(\mathcal{X})$.

The context algorithm yields, via the estimated transition probabilities in Step 4, an estimate

$$\hat{P}_n = \hat{P}_{n;VLMC} \quad (2.7)$$

which is consistent for all finite dimensional distributions of suitably regular processes P on $\mathcal{X}^{\mathbb{Z}}$. [The notation $\hat{P}_n$ is exclusively used for the estimated VLMC $\hat{P}_{n;VLMC}$ on $\mathcal{X}^{\mathbb{Z}}$]. They don't necessarily need to be a VLMC, see Bühlmann and Wyner (1999) and Ferrari (1999). The basis for these results is consistency of the context algorithm for finding the true underlying structure of the context function $c(.)$ from a general stationary process P. Generally $c(.)$ is of infinite order. We consider then the truncated context function

$$c_n(x_{-\infty}^0) = \begin{cases} c(x_{-\infty}^0) & \text{if } w = c(x_{-\infty}^0) \text{ has length } \mathrm{card}(w) \leq d_n \\ x_{-d_n+1}^0 & \text{if } \mathrm{card}(c(x_{-\infty}^0)) > d_n \end{cases}, \tag{2.8}$$

where $(d_n)_{n\in\mathbb{N}}$ is a sequence with $d_n \to \infty$ as $n \to \infty$. Under suitable conditions on the process P and the sequence $(d_n)_{n\in\mathbb{N}}$, Ferrari (1999) shows that the estimated context function

$$\hat{c}_n(.) \text{ from the context algorithm,} \tag{2.9}$$

satisfies

$$\mathbb{P}[\hat{c}_n(.) = c_n(.)] \to 1 \ (n \to \infty).$$

Thus, since d_n [from the truncated context function $c_n(.)$] increases, the context algorithm finds consistently the finite structure of $c(.)$, and it grows longer memory to infinite length for structures producing infinitely long values of $c(.)$ [having infinite depth in the context tree representation τ_c].

2.3 Definition and tuning the VLMC-sieve bootstrap

The construction of the VLMC-sieve bootstrap is now straightforward by simulating from $\hat{P}_n$ in (2.7),

$$X_1^*, \ldots, X_n^* \sim \hat{P}_n. \tag{2.10}$$

The sieve-bootstrapped estimator is based on the plug-in rule,

$$\hat{\theta}_n^* = h_n(X_1^*, \ldots, X_n^*), \tag{2.11}$$

with h_n the defining function of the estimator $\hat{\theta}_n = h_n(X_1, \ldots, X_n)$. The fact that the VLMC-sieve bootstrap employs the plug-in rule can be an advantage in practice. The block bootstrap [Künsch, 1989] instead should be generally used in conjunction with 'pre-blocking' the original data which in turn often requires redesigning the computation of $\hat{\theta}_n^*$, see Example 1.1 in Section 1. For the estimator in (2.1), the following representation then holds,

$$\hat{\theta}_n^* = T(\hat{P}_{n,emp}^{*(m)}),\ \hat{P}_{n,emp}^{*(m)}(x) = (n-m+1)^{-1} \sum_{t=1}^{n-m+1} 1_{[(X_t^*,\ldots,X_{t+m-1}^*)=x]}\ (x \in \mathcal{X}^m).$$

The sieve bootstrap sample in (2.10), sampled from $\hat{P}_n = \hat{P}_{n;\alpha}$ with context function $\hat{c}_n = \hat{c}_{n;\alpha}$, depends on the choice of the α% cutoff in (2.6). For sample sizes between 200 and 1000, cutoffs in the range of 95 to 99 % often work satisfactory. A more objective choice is given by the AIC criterion,

$$\hat{\alpha}_{AIC} = argmin_{\alpha}\{-2\ell(\hat{P}_{n;\alpha}) + 2\text{card}(\tau_{\hat{c}_{n;\alpha}})(\text{card}(\mathcal{X}) - 1)\}.$$

The advantage of this estimate is its relatively fast computation: the model selection problem boils down to a one-dimensional search on a grid for an optimal α.

The context algorithm and the VLMC-sieve bootstrap are implemented in the statistical computing language R, freely available from http://www.r-project.org/. More details are given in Appendix A.

3 Theoretical properties

We present a result showing that for variance estimation, the VLMC-sieve bootstrap has better rates of convergence than the block bootstrap [Künsch, 1989] if the underlying process has exponentially fast decaying strength of dependence as separation lags grow. One main assumption about the data generating process is

(A) The process $(X_t)_{t\in\mathbb{Z}} \sim P$ is stationary and geometrically α-mixing with $\alpha_P(k) \leq const.\rho^k$ for some $0 \leq \rho < 1$.

We first give an asymptotic justification of the VLMC-sieve bootstrap for estimators as defined in (2.1). As smoothness requirement for the real-valued functional T we assume Hadamard differentiability, cf. Van der Vaart and Wellner (1996, Ch.3.9.1).

Theorem 3.1 *Let $(X_t)_{t\in\mathbb{Z}} \sim P$ be a stationary $\mathcal{X}$-valued process satisfying (A) and (B1), (B2), (B3) given in Appendix B.1. Let $\hat{\theta}_n$ be as in (2.1) with T continuously Hadamard differentiable at $P^{(m)}$. Let $\hat{\theta}_n^*$ be its sieve-bootstrapped version defined in (2.11) and assume that [the existing limiting variance] $\lim_{n\to\infty} n\,\mathrm{Var}(\hat{\theta}_n)$ is strictly bigger than zero. Then*

$$\sup_{x\in\mathbb{R}} |\mathbb{P}^*[n^{1/2}(\hat{\theta}_n^* - \theta^*) \leq x] - \mathbb{P}[n^{1/2}(\hat{\theta}_n - \theta) \leq x]| = o_P(1),$$

where $\theta^* = T(\hat{P}_n^{(m)})$ with $\hat{P}_n^{(m)}$ from (2.7).

A proof is given in the Appendix B.2. An estimator which satisfies the Hadamard differentiability assumption is described in (2.2).

In the following we describe some accuracy results. We restrict the focus to estimators

$$\hat{\theta}_n = (n - m + 1)^{-1} \sum_{t=1}^{n-m+1} g(X_t, \ldots, X_{t+m-1}),\ (m \in \mathbb{N}) \tag{3.1}$$

where $g : \mathcal{X}^m \to \mathbb{R}$ is an arbitrary function.

Theorem 3.2 *Let $(X_t)_{t\in\mathbb{Z}} \sim P$ be a stationary $\mathcal{X}$-valued process, satisfying (A) and (B1'), (B2'), (B3') given in Appendix B.1. Let $\hat{\theta}_n$ be as in (3.1) and $\hat{\theta}_n^*$ its sieve-bootstrapped version from (2.11). Then,*

$$nVar^*(\hat{\theta}_n^*) - nVar(\hat{\theta}_n) = O_P(\log(n)^6 n^{-1/2}).$$

A proof is given in the Appendix B.2. The merit of Theorem 3.2 is an excellent rate of the variance estimator. By assumption, the underlying mixing coefficients [measuring strength of dependence] decay exponentially and the convergence rate $\log(n)^6 n^{-1/2}$ almost achieves the finite parametric bound. In contrast, the block bootstrap [Künsch, 1989] is not rate adaptive: under suitable regularity conditions,

$$\mathbb{E}\left[\left(nVar^{B*}(\hat{\theta}_n^{B*}) - nVar(\hat{\theta}_n)\right)^2\right] \sim const.n^{-2/3}, \tag{3.2}$$

achieved with the optimal blocklength parameter $\ell = const.n^{1/3}$, see Künsch (1989). Regardless on the decay of mixing coefficients, the bound $O_P(n^{-1/3})$ for the rate of the block bootstrap variance cannot be improved: it is clearly non-optimal whenever - roughly speaking - the true underlying process is exponentially mixing. Examples of such processes are suitably regular $\mathcal{X}$-valued Markov chains of high order, or processes $(X_t)_{t\in\mathbb{Z}}$ which are obtained from a scalar quantizer $q : \mathbb{R} \to \mathcal{X}$, $X_t = q(Y_t)$, where $(Y_t)_{t\in\mathbb{Z}}$ is an exponentially mixing $\mathbb{R}$-valued sequence.

Assumption (A) with the exponential decay of the mixing coefficient doesn't seem essential for consistency in Theorem 3.1. However, the context algorithm has only been theoretically analyzed using such an assumption [Ferrari, 1999]. Under a polynomial decay of the mixing coefficient in (A), the rate in Theorem 3.2 is expected to be slower: the superiority over the block bootstrap would then be less pronounced or even not true anymore.

4 Numerical examples

We analyze whether the asymptotic properties from Section 3 can be exploited for finite samples.

4.1 Variance estimation in simulations

The asymptotic differences in variance estimation described by Theorem 3.2 and formula (3.2) can be well seen in finite-sample problems, and the gain of the VLMC-sieve bootstrap is often substantial.

We consider the following binary, non-Markovian models. The first has a maximally sparse, infinitely long memory [represented by a binary context tree growing only one main branch with contexts $\{0, 1, 10, 11, 110, 111, 1110, 1111, \ldots\}$],

$$X_t = 1_{[Y_t>0]},\ Y_t = 0.8Y_{t-1}1_{[Y_{t-1}>0]} + \varepsilon_t,\ (t \in \mathbb{Z}), \tag{4.1}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d. innovation sequence, $\varepsilon_t \sim t_6$ independent from $\{Y_s; s < t\}$. The second has a non-sparse, infinitely long memory [represented by a full binary context tree],

$$X_t = 1_{[Y_t>0]},\ Y_t = 0.8Y_{t-1} + \varepsilon_t,\ (t \in \mathbb{Z}), \tag{4.2}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d. innovation sequence as in (4.1). The third model has also a non-sparse, infinitely long memory but only polynomial decay of dependence as lags between observations grow,

$$X_t = 1_{[Y_t>0]},\ Y_t = \sum_{j=1}^{\infty} \theta_j \varepsilon_{t-j} + \varepsilon_t,\ (t \in \mathbb{Z}),$$
$$\theta_j = 10/(j+10) 1_{[j \le 90]} + 90^{2.05}/(10 j^{2.05}) 1_{[j>90]} \tag{4.3}$$

where $(\varepsilon_t)_{t \in \mathbb{Z}}$ is an i.i.d. innovation sequence, $\varepsilon_t \sim \mathcal{N}(0,1)$ independent from $\{Y_s; s < t\}$. The polynomial decay of autocorrelations and mixing coefficients in model (4.3) is of the order $k^{-2.05}$ as separation lag k increases and thus summable. The fourth model has only dependence between values separated by at most 1 lag [so-called 1-dependence],

$$X_t = 1_{[Y_t>0]},\ Y_t = 10/11 \varepsilon_{t-1} + \varepsilon_t,\ (t \in \mathbb{Z}), \tag{4.4}$$

where $(\varepsilon_t)_{t \in \mathbb{Z}}$ is an i.i.d. innovation sequence as in (4.3); also, the first MA-coefficient is as in (4.3). These for models were chosen to represent different aspects: models (4.1) and (4.2) have geometrically decaying strength of dependence [e.g. mixing] and are falling in the range where the theory for the VLMC-sieve bootstrap as presented in Theorem 3.2 is worked out. It is interesting to see whether there are differences in performance for the sparse and the non-sparse, infinitely long memories [although the sparse model (4.1) exhibits also less strong dependence in terms of autocorrelations]. Model (4.3) is chosen with a decay of strength of dependence such that the mean square error rate $O(n^{-2/3})$ for variance estimation is close to optimal. It is thus a model for which the block bootstrap has asymptotically almost optimal rates of convergence for variance estimation. Model (4.4) represents one of the simplest extensions of an i.i.d. sequence: the block bootstrap should perform very well because already the fixed, non-increasing blocklength $\ell = 2$ yields the mean square error convergence rate of $O(n^{-1})$ for variance estimation. This is not a contradiction to (3.2) where we have assumed that the blocklength does increase with sample size which is necessary as soon as dependence has longer range than just over finitely many consecutive observations. The VLMC-sieve bootstrap is not tailored for such situations because the memory in (4.4) is still infinitely long, despite the simple structure of dependence.

Of interest are always the stationary binary processes $(X_t)_{t \in \mathbb{Z}}$ in (4.1)–(4.4). The sample size is $n = 512$. As a simple estimator, we consider $\hat{\theta} = \overline{X}_n = n^{-1} \sum_{t=1}^{n} X_t$.

The VLMC-sieve bootstrap is run with different cutoff tuning parameters in the context algorithm for constructing the estimate $\hat{P}_n = \hat{P}_{n;VLMC}$. For comparison, we apply the block bootstrap [Künsch, 1989] which needs specification of the so-called blocklength. We use here $\ell = 8 = n^{1/3}$ which is asymptotically of correct order [see formula (3.2)], and also an estimated blocklength $\hat{\ell}$ for block bootstrap variance estimation which has been proposed by Bühlmann and Künsch (1999).

The results are summarized in Figure 4.1 and Tables 4.1 and 4.2. Visual detection of significant differences in Figure 4.1 is not always obvious. Moreover, the comparison should be made on the basis of having different procedures on the same data realizations. Table 4.2 shows results from paired Wilcoxon tests. Denoting by $U_i = (n \operatorname{Var}^*(\overline{X}_n^*) - n \operatorname{Var}(\overline{X}_n))^2$ for the ith simulated data-set with a certain bootstrap method and by V_i the analogous quantity but with another bootstrap method [or other tuning parameter], the null-hypothesis is $\mathbb{E}[U_i - V_i] = 0$ and the alternative is two-sided. Test results as in

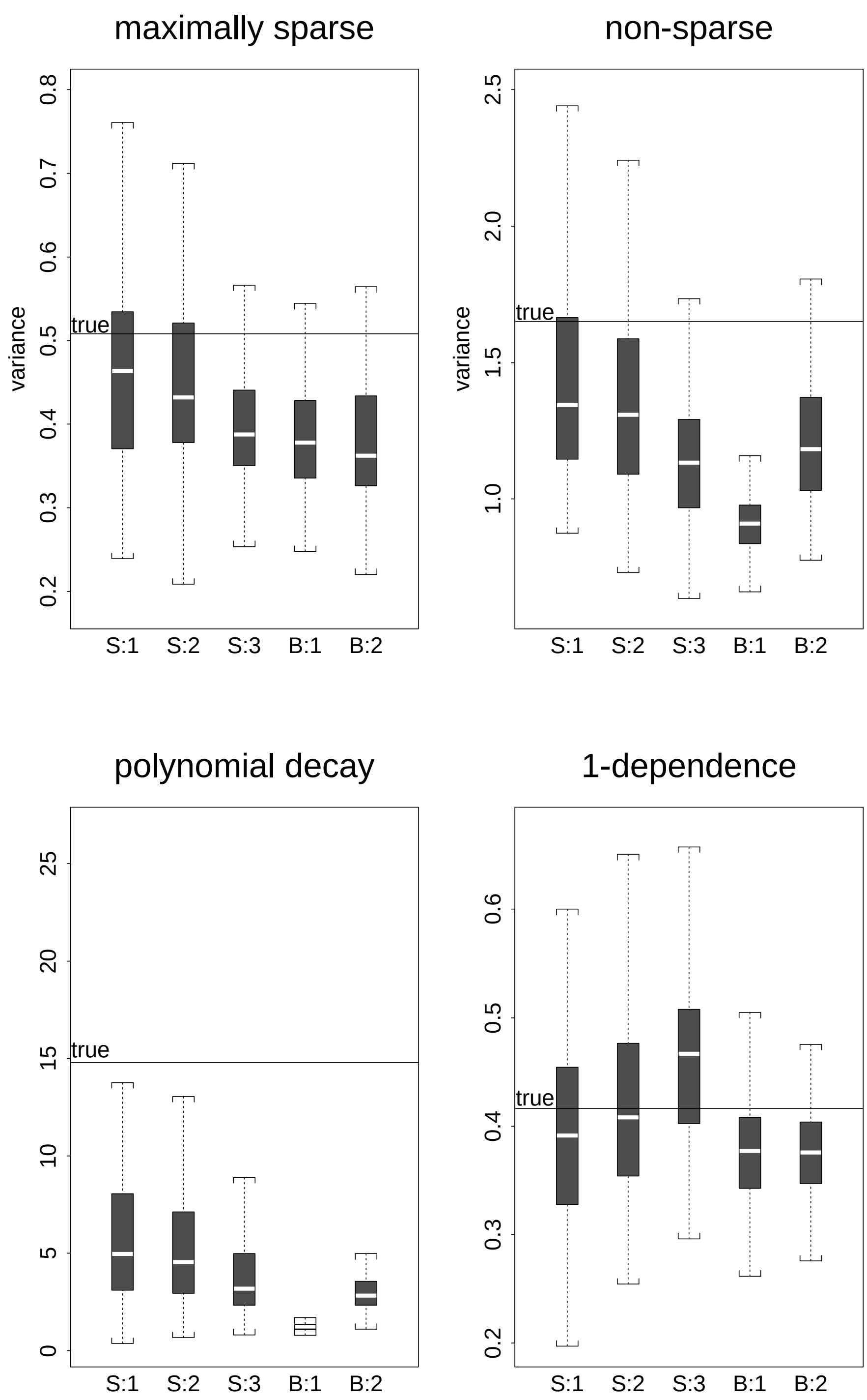

Figure 4.1: Bootstrap variance estimation of $n^{1/2}\overline{X}_n$ by $n\,\mathrm{Var}^*(\overline{X}_n^*)$. Sample size $n = 512$. Left panel: model (4.1); right panel: model (4.2). The target $n\,\mathrm{Var}(\overline{X}_n)$ is indicated by the horizontal lines. VLMC-sieve bootstraps are indicated by 'S:1', 'S:2', 'S:3' with 95%,97% and 99% cutoffs [see (2.6)], respectively; block bootstraps are indicated by 'B:1' with $\ell = 8$ and 'B:2' with estimated $\hat{\ell}$. 100 simulation runs, 500 bootstrap replicates per simulation run.

RMSE	VLMC, 95%	VLMC, 97%	VLMC, 99%	B, $\ell = 8$	B, $\hat{\ell}$
Sparse	0.068	0.059	0.069	0.076	0.087
Non-Sparse	0.077	0.079	0.114	0.201	0.095
Poly. decay	0.429	0.467	0.560	0.844	0.644
1-dependence	0.049	0.044	0.039	0.028	0.020

Table 4.1: Relative mean squared errors RMSE of bootstrap variance estimates $\mathbb{E}[(\mathrm{Var}^*(\overline{X}_n^*) - \mathrm{Var}(\overline{X}_n))^2]/\mathrm{Var}(\overline{X}_n)^2$. Sparse model (4.1) and non-sparse model (4.2). Sample size $n = 512$. VLMC-sieve bootstrap with 95%,97% and 99% cutoffs, see (2.6); block bootstrap with $\ell = 8$ and estimated $\hat{\ell}$. 100 simulation runs, 500 bootstrap replicates per simulation run.

Sparse	VLMC, 95%	VLMC, 97%	VLMC, 99%	B, $\ell = 8$	B, $\hat{\ell}$
VLMC, 95%	–	(−)0.562	(+)0.041	(+)0.001	(+)0.000
VLMC, 97%	–	–	(+)0.004	(+)0.000	(+)0.000
VLMC, 99%	–	–	–	(+)0.289	(+)0.008
B, $\ell = 8$	–	–	–	–	(+)0.007

Non-sparse	VLMC, 95%	VLMC, 97%	VLMC, 99%	B, $\ell = 8$	B, $\hat{\ell}$
VLMC, 95%	–	(+)0.333	(+)0.000	(+)0.000	(+)0.000
VLMC, 97%	–	–	(+)0.000	(+)0.000	(+)0.006
VLMC, 99%	–	–	–	(+)0.000	(−)0.011
B, $\ell = 8$	–	–	–	–	(−)0.000

Poly. decay	VLMC, 95%	VLMC, 97%	VLMC, 99%	B, $\ell = 8$	B, $\hat{\ell}$
VLMC, 95%	–	(+)0.000	(+)0.000	(+)0.000	(+)0.000
VLMC, 97%	–	–	(+)0.000	(+)0.000	(+)0.000
VLMC, 99%	–	–	–	(+)0.000	(+)0.000
B, $\ell = 8$	–	–	–	–	(−)0.000

1-dependence	VLMC, 95%	VLMC, 97%	VLMC, 99%	B, $\ell = 8$	B, $\hat{\ell}$
VLMC, 95%	–	(−)0.271	(−)0.576	(−)0.002	(−)0.000
VLMC, 97%	–	–	(−)0.669	(−)0.067	(−)0.002
VLMC, 99%	–	–	–	(−)0.123	(−)0.020
B, $\ell = 8$	–	–	–	–	(−)0.054

Table 4.2: P-values from paired Wilcoxon-tests for MSE of bootstrap variance estimates $n\,\mathrm{Var}^*(\overline{X}_n^*)$. On top: model (4.1); at bottom: model (4.2). Sample size $n = 512$. VLMC-sieve bootstrap 95%, 97% and 99% cutoffs, see (2.6); block bootstrap with $\ell = 8$ and estimated $\hat{\ell}$. The symbols (+) and (−) indicate in favor of the method in the row and in disfavor, respectively. 100 simulation runs, 500 bootstrap replicates per simulation run.

Table 4.2 are thus an appropriate tool for assessing significance to judge differences in performance between two methods.

It is fair to say that overall, the VLMC-sieve performed better than the block bootstrap. In models (4.1) and (4.2), the VLMC-sieve bootstrap with 95% or 97% cutoff [see (2.6)] is significantly better than any of the two block bootstraps. In the sparse model (4.1), the VLMC method is best [with respect to MSE] with the 97% cutoff $\chi^2_{1;0.97}/2$ and significantly outperforms the block bootstrap with $\ell = 8$, which is slightly better than with estimated $\hat{\ell}$. In the non-sparse model (4.2), the best VLMC-sieve bootstrap is with the 95% cutoff. It is better, with a high degree of significance, than the better of the two block methods, i.e. using the estimated $\hat{\ell}$. In model (4.3) with polynomial decay of strength of dependence, the block bootstrap performed poorly in comparison to the VLMC-sieve method. This comes a bit as a surprise because from asymptotic theory, the block bootstrap mean square error rate $O(n^{-2/3})$ [see (3.2)] is close to optimal. Finally, the block is better than the VLMC-sieve bootstrap in model (4.4) with 1-dependence: this is consistent with asymptotic theory and the fact that the block bootstrap is tailored, by construction, for processes which are m-dependent [with m small]. Generally, we found that the block bootstrap is a bit more sensitive to the specification of the blocklength than the VLMC-sieve technique to the choice of the cutoff tuning parameter.

Similar empirical results were found for the smaller sample size $n = 128$.

An alternative for variance estimation is via linearization of the estimator

$$\hat{\theta}_n \approx (n-m+1)^{-1} \sum_{t=1}^{n-m+1} \mathrm{IF}(X_t^{t+m-1}; P^{(m)})$$

with $\mathrm{IF}(\cdot; P)$ the influence function of the functional T in (2.1) at $P^{(m)}$, see Hampel et al. (1986, Ch.2.1). Thus,

$$\mathrm{Var}(\hat{\theta}_n) \sim n^{-1} \sum_{k=-\infty}^{\infty} R_{\mathrm{IF}}(k) = n^{-1} 2\pi f_{\mathrm{IF}}(0),$$

with $R_{\mathrm{IF}}(\cdot)$ and $f_{\mathrm{IF}}(\cdot)$ the autocovariance and spectral density of $(\mathrm{IF}(X_t^{t+m-1}; P))_{t\in\mathbb{Z}}$. This motivates the variance estimate

$$\widehat{\mathrm{Var}}(\hat{\theta}_n) = n^{-1} 2\pi \hat{f}_{\widehat{\mathrm{IF}}}(0),$$

where $\hat{f}_{\widehat{\mathrm{IF}}}(0)$ denotes a spectral density estimate at zero based on estimated influence functions $(\mathrm{IF}(X_t^{t+m-1}; \hat{P}_{n;emp}))_{t=1}^{n-m+1}$ with $\hat{P}_{n;emp}$ as in (2.1).

For the spectral density estimator at zero we use a lag-window estimators. One is with a C_2-window, twice continuously differentiable at zero and with automatic local bandwidth choice as in Bühlmann (1996). The other is with the trapezoid window (Politis and Romano, 1995) which asymptotically adapts to the underlying smoothness of the density: here, we have used the smoothing parameters which are optimal for the simulation. Of course, this is over-realistic and not a data-driven rule for a practical situation. From an asymptotic point of view, both estimates have good mean square error properties, namely $O(n^{-4/5})$ and $O(\log(n)n^{-1})$ for the C_2 and the trapezoid window, respectively [for geometric mixing processes]; these rates are only for spectral density estimation but

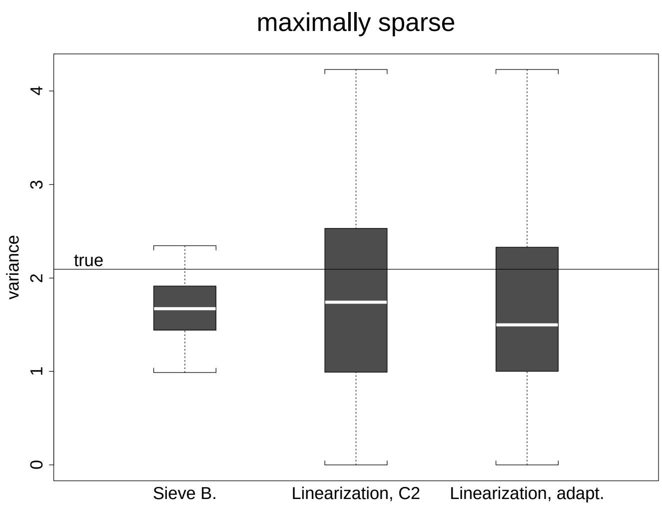

Figure 4.2: Variance estimation of $n^{1/2}N(11110)/N(1110)$ in sparse model (4.1) with sieve bootstrap and linearization using a C_2 window and the adaptive trapezoid window. Sample size $n = 128$. 100 simulation runs, 200 bootstrap replicates per simulation run.

don't reflect the effect of linearization. The methods are supposed to work well for 'nearly linear' estimators. But the linearization effect can cause substantial finite sample errors. Moreover, this alternative is less automatic, requiring explicit calculation of influence functions. Figure 4.2 displays results of such alternative variance estimates, compared to the sieve bootstrap using the 95% cutoff. The numerical study is with a very nonlinear estimator

$$\hat{\theta}_n = \frac{N(11110)}{N(1110)},$$

the model free estimate of $P(1|1110)$. The data is generated from the sparse model in (4.1) with sample size $n = 128$. The gain with the sieve bootstrap is impressive, because the strong nonlinearity of the estimator causes severe problems in the linearization method. This occurred without fine tuning of the VLMC bootstrap [fixed 95% cut-off], whereas the linearization with the adaptive trapezoid window was optimized for best mean squared error in the simulation study [not a data-driven rule].

The effect of nonlinearity in $\hat{\theta}_n$ with the very small sample size $n = 128$ for such an estimator is a rather extreme case: nevertheless, the VLMC-sieve bootstrap which doesn't rely explicitly on linearization [as any bootstrap method] performed very well.

4.2 Coverage of confidence intervals in simulations

We also examine construction of two-sided confidence interval with the estimator $\hat{\theta} = \overline{X}_n$ for $\theta = \mathbb{E}[X_t] = 0.725$ in model (4.1) and $\theta = 0.5$ in model (4.2) on nominal coverage level 0.9. The sample size is now chosen as $n = 128$.

We consider first order VLMC and block bootstrap intervals and corrections thereof. Due to the discreteness of observations X_t [latticeness of the problem], correction of confidence intervals doesn't seem worthwhile from an asymptotic point of view. However, there might be still some considerable gain for finite samples. For a related discussion see Hall (1987) and Woodroofe and Jhun (1989). For the VLMC method we examine the double bootstrap and bootstrapping the studentized mean. For the latter, we use a first level bootstrap to estimate $\mathrm{Var}(\overline{X}_n)^{1/2}$ and then a second level (double) bootstrap for the studentized statistic. The double bootstrap works as follows. Consider construction of a two-sided confidence interval which covers θ with nominal probability $1-\alpha$. A first order VLMC bootstrap interval is given by $[\hat{\theta}_n - \hat{r}_{1-\alpha/2}, \hat{\theta}_n - \hat{r}_{\alpha/2}]$, where

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

where $\theta^* = T(\hat{P}_n^{(m)})$ with T and $\hat{P}_n$ as in (2.1) and (2.7), respectively. Now consider an additive correction of the original nominal coverage level by using the double bootstrap. Based on $X_1^*, \ldots, X_n^*$, run the VLMC-sieve bootstrap to obtain $X_1^{**}, \ldots, X_n^{**} \sim \hat{P}_n^*$. Now,

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

where $\theta^{**} = T(\hat{P}_n^{*(m)})$. Put

$$\hat{a}(1-q) = \mathbb{P}^*[\hat{\theta}_n^* - \hat{r}_{1-q/2}^* \leq \theta^* \leq \hat{\theta}_n^* - \hat{r}_{q/2}^*],$$

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

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

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

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

as a two-sided, double bootstrap confidence interval for θ with nominal coverage level $1-\alpha$. For the block bootstrap we use a version of BC_a as given by Götze and Künsch (1996).

The tuning parameters of the bootstraps were chosen as follows: 95% cutoff for the VLMC and estimated $\hat{\ell}$ for the block bootstrap.

Coverage probabilities are given in Table 4.3. For the block bootstrap, the BC_a method increases performance a bit compared to the first order block intervals, although the difference is not relevant in the sparse model (4.1) and only weakly relevant in the non-sparse model (4.2). The first order VLMC shows remarkable gain over the first order block bootstrap in model (4.1), whereas there is no relevant difference in model (4.2). Correction of the first order VLMC bootstrap is worthwhile with the t-VLMC [using studentization] and the double VLMC bootstrap. Overall, the double VLMC bootstrap performs best in this small study.

Sparse	VLMC, 95%	Block, $\hat{\ell}$	Double VLMC, 95%	t-VLMC, 95%	Block BC$_a$, $\hat{\ell}$
coverage	0.81	0.75	0.88	0.98	0.78
median(length)	0.178	0.164	0.215	0.189	0.164
MAD(length)	0.034	0.023	0.065	0.065	0.022

Non-sparse	VLMC, 95%	Block, $\hat{\ell}$	Double VLMC, 95%	t-VLMC, 95%	Block BC$_a$, $\hat{\ell}$
coverage	0.74	0.71	0.84	0.85	0.75
median(length)	0.276	0.262	0.368	0.322	0.258
MAD(length)	0.063	0.041	0.122	0.130	0.041

Table 4.3: Coverage probabilities for two-sided confidence interval on nominal 90% level with median and mean absolute deviation (MAD) of their lengths. Sample size $n = 128$. On top: model (4.1); at bottom: model (4.2). First order accurate VLMC-sieve and block bootstrap intervals; and corrections with the double VLMC-sieve, the VLMC-sieve for studentized statistic [t-VLMC] and the BC$_a$ block bootstrap. Based on 100 simulations and 500 bootstrap replicates.

4.3 DNA sequence

We analyze here homogeneity of a DNA sequence with base elements in $\mathcal{X} = \{A, C, G, T\}$ by searching for so-called $C\mathbf{p}G$-islands [Bird, 1987]. The chance for occurrence of the pair of consecutive base elements (C, G) may not be constant over the whole DNA string and stretches with high chances are classified as interesting potential starting regions of a gene.

Our data began as a single 100-thousand base contiguous stretch of DNA from the Drosophila genome (Genbank number DS02740). The sequence is then segmented into genes (which code for amino acid sequences) and non-genes which are assumed to be ignored by the cell chemistry. Physically, the genes are spaced apart and separated by non-genes. Moreover, the genes are further segmented into alternating stretches of coding regions called exons and non-coding regions called introns. The working data is then a single sequence of exons by concatenating all the exons (in the given order): it is an $\mathcal{X}$-valued sequence of length $n = 25460$ with $\mathcal{X} = \{A, C, G, T\}$.

The curve of smoothed frequencies of CG from a triangular window with 1000 values is

$$\hat{\theta}_n(t) = \sum_{j=t-499}^{t+499} w_j(t) 1_{[X_j = C, X_{j+1} = G]},$$

$$w_j(t) = h_j(t) / \sum_{j=t-499}^{t+499} h_j(t),\ h_j(t) = \max\{1 - |t-j|/500, 0\}\ (t = 500, \ldots, n-500).$$

The z-score statistic is defined as,

$$\hat{\zeta}_n(t) = \frac{\hat{\theta}_n(t) - \theta^*}{(\mathrm{Var}^*(\hat{\theta}^*_n(t)))^{1/2}}\ (t = 500, \ldots, n-500),$$

where $\theta^* = \mathbb{P}_{\hat{P}_n}[X_t = C, X_{t+1} = G]$ is the chance for CG in a fitted VLMC $\hat{P}_n$ [θ^* is evaluated as empirical frequency from a very long realization from $\hat{P}_n$]. Note that

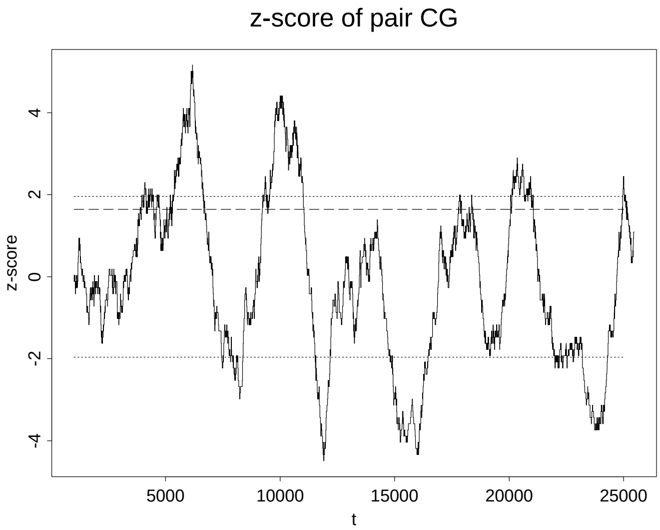

Figure 4.3: Z-scores $\hat{\zeta}_n(t)$ of pair CG indicated by solid line. One- and two-sided acceptance regions on 5% significance level indicated by dashed and dotted lines, respectively [one-sided against alternatives leading to large z-scores].

$\theta^* = \mathbb{E}^*[\hat{\theta}_n^*(t)]$ for all t and $\mathrm{Var}^*(\hat{\theta}_n^*(t))$ is constant over t since $\hat{P}_n$ is stationary. Such z-scores, for the whole data sequence rather than within a window, have been mainly proposed under an estimated full Markov model, cf. Schbath et al. (1995), Leung et al. (1996) and Schbath (1997) [note that here, $\mathrm{Var}(\hat{\theta}_n(t) - \theta^*) = \mathrm{Var}(\hat{\theta}_n(t))(1 + o(1))$ which explains the denominator in $\hat{\zeta}_n(t)$]. There, the aim is to find extraordinary words [relative to a full Markov model] rather than homogeneity classification: the task does not involve the 'best' model describing the data. In contrast, for homogeneity classification with our z-scores $\hat{\zeta}_n(t)$, the choice of the non-flexible full Markov model seems too restrictive: either the full Markov model is of low order so that z-scores only lead to indications relative to an over-simplified model; or the full Markov model has high order so that the variance estimate $\mathrm{Var}^*(\hat{\theta}_n^*(t))$ [and θ^*] may be extremely poor. Our z-scores relative to a flexible VLMC model thus allow more realistic indications.

The general testing hypothesis is that $\mathbb{P}_P[X_t = C, X_{t+1} = G] = \theta(t) \equiv \theta$ is constant over all $t \in \mathbb{Z}$ and the underlying process $(X_t)_{t\in\mathbb{Z}} \sim P$ is stationary: the alternative is its complement. Under the hypothesis, the nuisance parameter P is consistently estimated by an approximating VLMC, denoted by $\hat{P}_n$. Then, under the hypothesis, the z-scores $\hat{\zeta}_n(t)$ are asymptotically $\mathcal{N}(0,1)$-distributed and standard normal quantiles are used for constructing acceptance regions.

One- and two-sided acceptance regions for the z-scores on the 5% significance level are displayed in Figure 4.3. A violation at time t indicates that the process $(X_s)_{s=t-499}^{t+499}$ is significantly different from being stationary; but not necessarily due to non-constant $\theta(t)$. Potential locations of $C\mathbf{p}G$ islands are easily visible and may be worthwhile to analyze with biological experiments.

A different strategy for segmentation of DNA has been pursued with Hidden Markov models, cf. Churchill (1989) and Muri (1998). There, the stochastic modeling philosophy is somewhat different. Hidden Markov models are usually based on locally simple structures but build up complexity by allowing switching between regimes. Our approach with VLMC's is without explicit switching regimes but with potentially high global complexity.

5 Conclusions

We propose the VLMC-sieve bootstrap for stationary categorical time series. An alternative procedure is the block bootstrap which is designed for more general situations. Analytical techniques based on linearization can be very inefficient when the estimator of interest is highly nonlinear, see Section 4.1.

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

In the specific, important niche of categorical time series, the VLMC-sieve bootstrap outperforms the block scheme, provided the process has sufficiently fast decaying mixing structure: this is asymptotically described by Theorem 3.2 for variance estimation and it is also empirically found for finite sample cases. The practical relevance of this superiority applies to many real situations, where dependence in the data is far from long memory.

Another advantage of the VLMC-sieve bootstrap is the construction of bootstrapped estimators with the plug-in rule, see formula (2.11): the method thus shares the computational convenience of Efron's (1979) bootstrap. In contrast, block bootstrapped estimators are generally not defined via the plug-in rule and the computation of $\hat{\theta}^{*\text{block boot}}$ may need to be redesigned which can become an awkward task, see Example 1.1 in Section 1. The sophisticated resampling with the tree structured context algorithm for fitting a VLMC, which is a drawback in practice, can be efficiently implemented with $O(n\log(n))$ essential operations and is now available in R, see Appendix A.

A Appendix: Computation and software

The context algorithm and the VLMC-sieve bootstrap are implemented in the statistical computing language R, freely available from http://www.r-project.org/. The package "VLMC" should be downloaded. The exact commands in R look as follows:

> **install.packages("VLMC")**

> **library(VLMC)**

> **fit <- vlmc(series,cutoff=C)**

This is the VLMC fit of the data with the context algorithm. The default **cutoff** value is C= $\chi^2_{\text{card}(\mathcal{X})-1;0.95}/2$.

> **simulate.vlmc(fit,resample size,integer.return=F)**

This is the VLMC-sieve bootstrap sample. The default value is **integer.return**=F yielding letters from the alphabet $\mathcal{X}$.

For tuning the cutoff value, we search with the AIC statistic. The log-likelihood $\ell(\hat{P}_n)$ of a fitted VLMC is computed by

> **entropy(fit)**

The cardinality of the estimated context tree is given by
> **fit$size["context"]**
The alphabet size is automatically given by
> **fit$alpha.len**
The AIC statistic is thus given by
> $-2*$**entropy(fit)** + $2*$**(fit$size["context"])**$*$**(fit$alpha.len - 1)**
More on-line help is implemented in *R* with library(VLMC).

B Appendix: Theory

B.1 VLMC-sieve and assumptions

The stationary process $(X_t)_{t\in\mathbb{Z}}$ has a distribution P on $\mathcal{X}^{\mathbb{Z}}$. Estimation of P can be viewed two-fold. First, a sieve-approximation of P with a VLMC, which we denote by P_n; and secondly, estimation of P_n, denoted by $\hat{P}_n$.

The sequence of VLMC sieve-approximations P_n is defined as

$$\begin{aligned}&P_n \text{ is a VLMC with context function } c_n(.),\\ &\text{and with transition probabilities } P_n(x|w) = \begin{cases} P(xw)/P(w) & \text{if } P(w) \neq 0 \\ 0 & \text{otherwise}\end{cases},\\ &\text{for } w \in \tau_{c_n}, \text{ where } c_n(.) \text{ is defined in (2.8).}\end{aligned} \tag{B.1}$$

We denote by p_n the order of $c_n(.)$ and consequently of the VLMC P_n [which can be embedded by a Markov chain of order p_n]. Note that $p_n \leq d_n$ [the truncation of $c_n(.)$], and strictly smaller than d_n for all n sufficiently large only if P is Markovian of finite order. As we will assume in (B1) and (B2) below, $d_n = const. \log(n)$ so that $p_n = O(\log(n))$. Aspects of the error between the true underlying process P and its sieve approximation P_n are given by the following dependence coefficients

$$\nu(k) = \nu_P(k) = \sup_{x_{-\infty}^1 \in \mathcal{X}^\infty} |P(x_1|x_{-k+1}^0) - P(x_1|x_{-\infty}^0)| \; (k \in \mathbb{N}). \tag{B.2}$$

The coefficient $\nu(k)$ is different from the more familiar notions of mixing [see Doukhan, 1994]. It describes the remaining dependence after conditioning on the previous k lagged values; the mixing notion instead describes the dependence of events being separated by k lags. The ν-dependence measure in (B.2) is the right notion for describing the quality of a Markov-sieve approximation.

Remark B.1. If P is Markovian of possibly high but finite order $p < \infty$, then $P_n \equiv P$ for all n sufficiently large, due to the properties of $c_n(.)$. Also, $\nu_P(k) = 0$ for all $k \geq p$.

Estimation of P_n is based on the context algorithm,

$$\begin{aligned}&\hat{P}_n \text{ with context function } \hat{c}_n(.),\\ &\text{and transition probabilities } \hat{P}_n(x|w) = N(xw)/N(w) \text{ for all } w \in \tau_{\hat{c}_n},\end{aligned} \tag{B.3}$$

where $\hat{c}_n(\cdot)$ is as in (2.9).

Besides assumption (A) from Section (3), the following conditions guarantee good behavior of $\hat{P}_n$ in (B.3) and in turn also for the VLMC-sieve bootstrap in (2.11).

(B1) The sieve approximations P_n as defined in (B.1), having context functions $c_n(.)$ with truncation d_n as defined in (2.8), satisfy: For some $d_n \sim C_d \log(n)$ with $0 < C_d < \infty$,

$$\text{card}(\tau_{c_n}) = o(n),$$
$$p_{min,T}^{-1} = O(n^{\beta}) \text{ for some } 0 \le \beta < 1/2, \text{ where } p_{min,T} = \min_{w \in \tau_{c_n}^T} P(w),$$
$$p_{sep,T}^{-2} = O\left(\frac{(np_{min,T}^{(1-\gamma)/2})^{1-\gamma}}{\log(n)^{1+\delta}}\right) \text{ for some } 0 \le \gamma < 1,\ \delta > 0,$$
$$\text{where } p_{sep,T} = \min_{wu \in \tau_{c_n}^T, u \in \mathcal{X}} \sum_{x \in \mathcal{X}} |P(x|wu) - P(x|w)|,$$
$$\left(\min_{x \in \mathcal{X}, w \in \tau_{c_n}^T} P(x|w)\right)^{-1} = O(n),$$
$$\min_{x_{-n+2}^1 \in \mathcal{X}^n, w = x^0_{-\text{card}(w)+1} \in \tau_{c_n}} \frac{P(x_1|w)}{P(x_1|x^0_{-n+2})} \ge 1 - a_n \log(n) n^{-1} \text{ for some sequence } a_n \to 0.$$

(B2) For some $d_n \sim C_d \log(n)$ with $0 < C_d < \infty$, $p_{min,p_n}^{-1} = O(n^{\lambda})$ $(0 < \lambda < \infty)$, where $p_{min,p_n} = \min_{x \in \mathcal{X}^{p_n}} P(x)$ and p_n the order of $c_n(.)$.

(B3) $\nu_P(k) \le const.k^{-1+\iota}$ $(\iota > 0)$ for all $k \in \mathbb{N}$.

Assumption (B1), together with assumption (A) from Section 3, ensures consistent estimation of $c_n(.)$ with the context algorithm, see Ferrari (1999). (B1) may be hard to check for given processes: an important example satisfying (B1) is given in Remark B.2 below. (B2) is a weak condition since λ can be much larger than 1; (B3) is a mild requirement on the decay of the dependence.

Moreover, we sometimes strengthen part of (B1) and (B3) to

(B1') As assumption (B1) but: $1 \le 1/(2\log(1/\xi)) \le C_d < \infty$ with ξ as in (B3') below; and $p_{min,T}^{-1} = O(p_n)$ [p_n the order of $c_n(.)$].

(B2') As assumption (B2) but C_d as in (B1').

(B3') $\nu_P(k) \le const.\xi^k$ for some $0 \le \xi < 1$.

The condition (B1') can be interpreted as a sparseness condition implying that $\text{card}(\tau_{c_n}^T) \le p_{min,T}^{-1} = O(p_n)$ is allowed to grow with the same order as p_n. This holds for trees $\tau_{c_n}^T$ being potentially non-sparse at first generations [nodes on top with small depths] but which grow only finitely many sparse branches to infinity as n increases. The example in (4.1) satisfies this condition.

Remark B.2. Consider P being Markovian [with or without variable length memory] of possibly high but finite order $p < \infty$, satisfying $\min_{x \in \mathcal{X}, w \in \tau_c} P(x|w) > 0$. Then, assumption (A) from Section 3, (B1'), (B2') and (B3') hold.

B.2 Proofs

We first give an auxiliary result about convergence of relative frequencies.

Lemma B.1 *Assume that $(X_t)_{t\in\mathbb{Z}} \sim P$ satisfies (A). Let $\mathcal{A}$ be a set of finite sequences $w = x^0_{-card(w)+1}$ and let $\ell_{\mathcal{A}} \geq \max\{card(w); w \in \mathcal{A}\}$ be an upper bound for the length of the elements in $\mathcal{A}$: $\mathcal{A}$ and $\ell_{\mathcal{A}}$ are allowed to depend on n. Assume that $\ell_{\mathcal{A}} = o(n)$ and $card(\mathcal{A}) = o(n^r)$ for some $0 < r < \infty$. Then,*

$$\sup_{w\in\mathcal{A}} |n^{-1}N(w) - P(w)| = \log(card(\mathcal{A}))^{1/2} O_P\left(\max\{\log(n)^{1/2}n^{-1/2}, \ell_{\mathcal{A}}^{1/2}n^{-1/2}\}\right)\ (n \to \infty).$$

Proof: The bias is bounded as

$$\mathbb{E}[n^{-1}N(w)] - P(w) = O(\mathrm{card}(w)/n) = O(\ell_{\mathcal{A}}/n).$$

Denote by $a_n = \log(\mathrm{card}(\mathcal{A}))^{1/2} \max\{\log(n)^{1/2}n^{-1/2}, \ell_{\mathcal{A}}^{1/2}n^{-1/2}\}$. Then, for any $\varepsilon > 0$, the stochastic part can be bounded as

$$\begin{aligned} &\mathbb{P}[a_n^{-1}|n^{-1}N(w) - \mathbb{E}[n^{-1}N(w)]| > \varepsilon] \\ \leq\ & const. \exp(-\varepsilon^2 a_n^2 q/8) + const.(\varepsilon a_n)^{-1/2} q \exp(\log(\rho)[(n - \ell_{\mathcal{A}})/(2q)]), \end{aligned}$$

where q is an integer in $[1, n/2]$ and the constants indicated by '*const.*' do not depend on w, see Bosq (1996, Ch.1.4, Th.1.3). By choosing $q = \min\{[C_q n/\log(n)], [C_q n/\ell_{\mathcal{A}}]\}$ for some suitable constant $0 < C_q < \infty$, we then obtain

$$\begin{aligned} &\mathbb{P}[a_n^{-1}|n^{-1}N(w) - \mathbb{E}[n^{-1}N(w)]| > \varepsilon] \\ \leq\ & const.\mathrm{card}(\mathcal{A})^{-C_q\varepsilon^2/8} + const.\varepsilon^{-1/2}n^{5/4}\exp(\log(\rho)\log(n)/(4C_q)). \end{aligned} \tag{B.4}$$

By choosing first $0 < C_q < \infty$ sufficiently small we obtain

$$\mathrm{card}(\mathcal{A})n^{5/4}\exp(\log(\rho)\log(n)/(4C_q)) \to 0\ (n \to \infty), \tag{B.5}$$

due to the assumption $\mathrm{card}(\mathcal{A})/n^r \to 0$ for some $0 < r < \infty$. Now observe that

$$\mathrm{card}(\mathcal{A})\mathrm{card}(\mathcal{A})^{-C_q\varepsilon^2/8} \to 0\ (\varepsilon \to \infty). \tag{B.6}$$

Therefore, by (B.4)-(B.6) we have

$$\sum_{w\in\mathcal{A}} \mathbb{P}[a_n^{-1}|n^{-1}N(w) - \mathbb{E}[n^{-1}N(w)]| > \varepsilon] \to 0\ (\varepsilon \to \infty,\ n \to \infty),$$

which completes the proof. □

We now prove that the sieve approximation P_n from (B.1) and its estimated version from (B.3) satisfy a Doeblin-type condition. Let

$$P_n^{(r)}(A, x) = \mathbb{P}_{P_n}[X^r_{r-p_n+1} \in A | X^0_{-p_n+1} = x]\ (A \subseteq \mathcal{X}^{p_n},\ x \in \mathcal{X}^{p_n})$$

be the r-step transition probability for the embedded Markov chain [of order p_n] driven by P_n.

Proposition B.1 *Assume that $(X_t)_{t\in\mathbb{Z}} \sim P$ satisfies (A), (B2) and (B3). Let $r_n = C_r \log(n)$ with $1 < C_r < \infty$ sufficiently large. Then, there exists an $n_0 \in \mathbb{N}$ such that*

$$\sup_{n\ge n_0} \sup_{x,y\in\mathcal{X}^{p_n};A\subseteq\mathcal{X}^{p_n}} |P_n^{(r_n)}(A,x) - P_n^{(r_n)}(A,y)| < 1-\kappa \text{ for some } 0<\kappa<1.$$

Proof: Write $Y_t = X_{t-p_n+1}^t$.

Step 1: Consider for any x, y and A,

$$\begin{aligned} & |P^{(r_n)}(A,x) - P^{(r_n)}(A,y)| = |\mathbb{P}_P[Y_{r_n} \in A|Y_0 = x] - \mathbb{P}_P[Y_{r_n} \in A|Y_0 = y]| \\ = \ & |\mathbb{P}_P[Y_{r_n} \in A|Y_0 = x] - P(A) - (\mathbb{P}_P[Y_{r_n} \in A|Y_0 = y] - P(A))| \\ \le \ & 2\alpha_P(r_n - p_n + 1)p_{min,p_n}^{-1}. \end{aligned}$$

By using (A), (B2) and choosing C_r sufficiently large [note that $p_n \le d_n = C_d \log(n) = O(\log(n))$], this implies that there exists an n_0 with

$$\sup_{n\ge n_0} \sup_{x,y,A} |P^{(r_n)}(A,x) - P^{(r_n)}(A,y)| < 1-\kappa$$

for some $0 < \kappa < 1$.

Step 2. For any x and A, consider the difference $|P_n^{(r_n)}(A,x) - P^{(r_n)}(A,x)|$. Denote by $x_0^{p_n-1} = x$. Then,

$$P^{(r_n)}(A,x) = \sum_{x_{r_n-p_n+1}^{r_n}\in A} \ \sum_{x_{p_n}^{r_n-p_n}\in\mathcal{X}^{r_n-1}} P(x_{r_n}|x_0^{r_n-1})P(x_{r_n-1}|x_0^{r_n-2})\cdots P(x_{p_n}|x_0^{p_n-1}). \quad \text{(B.7)}$$

And the same expansion also holds for $P_n^{(r_n)}(A,x)$ in terms of conditional probabilities $P_n(x_t|x_0^{t-1})$ $(p_n \le t \le r_n)$. Define the function

$$c^{(ext)}(x) = \begin{cases} x, & \text{if } xw \in \tau_c \text{ for some } w \in \cup_{m=1}^{\infty}\mathcal{X}^m \\ c(x), & \text{otherwise} \end{cases}$$

For the terms $P(x_t|x_0^{t-1})$ and $P_n(x_t|x_0^{t-1})$ $(p_n \le t \le r_n)$, the following cases are to be distinguished: (I) $c^{(ext)}(x_0^{t-1}) = c_n^{(ext)}(x_0^{t-1})$ for all n sufficiently large; (II) $\mathrm{card}(c^{(ext)}(x_0^{t-1})) > \mathrm{card}(c_n^{(ext)}(x_0^{t-1}))$ for all n sufficiently large; (III) $\mathrm{card}(c^{(ext)}(x_0^{t-1})) < \mathrm{card}(c_n^{(ext)}(x_0^{t-1}))$ for all n sufficiently large. Case (III) is impossible by definition of $c_n(.)$. In case (I), $P(x_t|c^{(ext)}(x_0^{t-1})) = P_n(x_t|c_n^{(ext)}(x_0^{t-1}))$ by definition of P_n. In case (II), $P(x_t|c^{(ext)}(x_0^{t-1})) = P_n(x_t|c_n^{(ext)}(x_0^{t-1})) + \Delta(x_0^t)$ with $\sup_{t,x_0^t} |\Delta(x_0^t)| \le \nu(p_n)$. Now using that $\sum_{x_t} P(x_t|x_0^{t-1}) = 1$ for all x_0^{t-1}, we obtain

$$\sup_{A,x} |P^{(r_n)}(A,x) - P_n^{(r_n)}(A,x)| \le \sum_{j=1}^{r_n} \binom{r_n}{j} \nu(p_n)^j \mathrm{card}(\mathcal{X})^j \le \sum_{j=1}^{r_n} (\nu(p_n) r_n \mathrm{card}(\mathcal{X}))^j,$$

which tends to zero by (B3) [either $p_n = d_n = C_d \log(n)$ or $\nu(p_n) \equiv 0$ for all n sufficiently large]. Together with Step 1 we complete the proof. □

We discuss next an analogue for the estimated VLMC sieve approximation $\hat{P}_n$.

Proposition B.2 *Assume that $(X_t)_{t\in\mathbb{Z}} \sim P$ satisfies (A) and (B1), (B2), (B3). Let $r_n = C_r \log(n)$ with $1 < C_r < \infty$ sufficiently large. Then, there exists an $n_0 \in \mathbb{N}$ and a sequence of sets B_n with $\mathbb{P}[B_n] \to 1$ such that*

$$\sup_{n\geq n_0} \sup_{x,y\in\mathcal{X}^{p_n}, A\subseteq\mathcal{X}^{p_n}} |\hat{P}_n^{(r_n)}(A,x) - \hat{P}_n^{(r_n)}(A,y)| < 1-\kappa \text{ on the set } B_n,$$
$$\hat{P}_n \text{ is } \phi\text{-mixing with } \phi_{\hat{P}_n}(k) < (1-\kappa)^{k/r_n} \text{ for } k \geq r_n, \text{ and on the set } B_n.$$

Proof: Denote by $\hat{c}_n(.)$ the estimated context function. The results in Ferrari (1999) show that there exists a sequence of sets D_n with $\mathbb{P}[D_n] \to 1$ such that $\hat{c}_n(.) = c_n(.)$ on D_n. Thus, on D_n, we can restrict ourselves to a version $\tilde{P}_n$ of $\hat{P}_n$, which have context functions $c_n(.)$ instead of $\hat{c}_n(.)$. Denote by

$$\Delta = \sup_{x\in\mathcal{X}, w\in\tau_{c_n}} |N(xw)/N(w) - P(x|w)|.$$

Using a first order Taylor expansion we obtain

$$|\Delta| \leq p_{min}^{-1} \sup_{x\in\mathcal{X}, w\in\tau_{c_n}} \left(|n^{-1}N(xw) - P(xw)| + |n^{-1}N(w) - P(w)|\right),$$

where $p_{min} = \min_{w\in\tau_{c_n}} P(w)$ satisfies $p_{min}^{-1} \leq p_{min,T}^{-1} = O(n^{\beta})$, due to (B1). Applying Lemma B.1 and using $\mathrm{card}(\tau_{c_n}) = o(n)$ from (B1) we obtain,

$$\sup_{x\in\mathcal{X}, w\in\tau_{c_n}} |n^{-1}N(xw) - P(xw)| = o_P(\log(n)n^{-1/2}).$$

And likewise for $\sup_{w\in\tau_{c_n}} |n^{-1}N(w) - P(w)|$. Hence,

$$\Delta = o_P(p_{min}^{-1} \log(n)n^{-1/2}) = o_P(n^{\beta-1/2}\log(n)). \tag{B.8}$$

We then analyze $\tilde{P}_n^{(r_n)}(A,x) - P_n^{(r_n)}(A,x)$ in the same way as in Step 2 from the proof of Proposition B.1: the bound becomes

$$\sup_{x,A} |\tilde{P}_n^{(r_n)}(A,x) - P_n^{(r_n)}(A,x)| \leq \sum_{j=1}^{r_n} (\Delta r_n \mathrm{card}(\mathcal{X}))^j = o_P(1),$$

since $\beta < 1/2$ from (B1) implies $\Delta r_n = o_P(n^{\beta-1/2}\log(n)^2) = o_P(1)$. Therefore, using the result from Proposition B.1, we can find a sequence of sets B_n [which are subsets of D_n] with $\mathbb{P}[B_n] \to 1$ on which the first assertion holds. The second assertion is an immediate consequence of the first. □

Proof of Theorem 3.1.
Using Hadamard differentiability of T at $P^{(m)}$, it is by now well known that it suffices to show

$$n^{1/2}(\hat{P}_{n,emp}^{(m)} - P^{(m)}) \Rightarrow W, \tag{B.9}$$
$$n^{1/2}(\hat{P}_{n,emp}^{*(m)} - P^{*(m)}) \Rightarrow W \text{ in probability}, \tag{B.10}$$

where '⇒' denotes weak convergence in $\mathcal{X}^m$ and W is a Gaussian random vector indexed by $\mathcal{X}^m$ [for the bootstrap, use continuity of the derivative T at $P^{(m)}$ and that $P^{*(m)}$ converges weakly to $P^{(m)}$]. Formula (B.9) follows by (A). For (B.10), we first show convergence of covariances

$$n\,\mathrm{Cov}^*(\hat{P}^{*(m)}_{n,emp}(x),\hat{P}^{*(m)}_{n,emp}(y)) - n\,\mathrm{Cov}(\hat{P}_{n,emp}(x),\hat{P}_{n,emp}(y)) = o_P(1),\ \ x,y\in\mathcal{X}^m. \quad \text{(B.11)}$$

This can be shown using the geometric mixing property of $\hat{P}_n$ [given in Proposition B.2], analogously to the proof of Theorem 3.2 [which gives a good convergence rate but needs more assumptions on $p_{min,T}$]. Finally, weak convergence in (B.10) to a Gaussian limit follows with the Bernstein technique: the 'small' blocks should have growing length of order not too big but larger than r_n, so that the geometric mixing property of $\hat{P}_n$, as stated in Proposition B.2, is sufficient to treat them as negligible. □

Proof of Theorem 3.2.
Choose $r_n = C_r\log(n)$ as in Proposition B.2 with $1 < C_r < \infty$ sufficiently large. Then set $b_n = [C_b\log(n)r_n]$ with $C_b \geq -1/(2\log(1-\kappa))$. Under the geometric mixing condition for P [see assumption (A)], the variance of $\hat{\theta}_n$ can be written as

$$n\,\mathrm{Var}(\hat{\theta}_n) = \sum_{k=-b_n}^{b_n} R_g(k) + O(n^{-1/2}), \quad \text{(B.12)}$$

where $R_g(k) = \mathrm{Cov}(g(X_0^{m-1}), g(X_k^{k+m-1}))$, cf. Doukhan (1994, Ch.1.2.2). Likewise, using the geometric mixing condition for the bootstrap $\hat{P}_n$ as stated in Proposition B.2,

$$n\,\mathrm{Var}^*(\hat{\theta}^*_n) = \sum_{k=-b_n}^{b_n} \hat{R}_g(k) + O_P(\log(n)n^{-1/2}), \quad \text{(B.13)}$$

where $\hat{R}_g(k) = \mathrm{Cov}^*(g((X^*)_0^{m-1}), g((X^*)_k^{k+m-1}))$. Formula (B.13) can be seen to hold as follows: on the set B_n [as stated in Proposition B.2],

$$\sum_{k=b_n+1}^{n-1} |\,\mathrm{Cov}^*(g((X^*)_0^{m-1}), g((X^*)_k^{k+m-1}))| \leq const. \sum_{k=b_n+1}^{\infty} (1-\kappa)^{k/r_n}.$$

Due to the choice of C_b and r_n, this yields the bound $O_P(n^{-1/2}\log(n))$. By formulae(B.12) and (B.13), it remains to analyze

$$\max_{0\leq k\leq b_n} |\hat{R}_g(k) - R_g(k)|.$$

Write

$$\hat{R}_g(k) - R_g(k) = \mathbb{E}^*[g((X^*)_0^{m-1})g((X^*)_k^{k+m-1})] - \mathbb{E}[g(X_0^{m-1})g(X_k^{k+m-1})] + O_P(n^{-1/2}),$$

where the O_P-term is uniform over k, arising from the difference $\mathbb{E}^*[g(X^*)_0^m]^2 - \mathbb{E}[g(X_0^m)]^2$ [since the m-dimensional marginal distribution can be estimated with convergence rate $1/\sqrt{n}$ and g is a bounded function due to the finiteness of $\mathcal{X}^m$]. Now,

$$\mathbb{E}^*[g((X^*)_0^m)g((X^*)_k^{k+m-1})] = \sum_{x_0^m, x_k^{k+m-1}\in\mathcal{X}^m} g(x_0^{m-1})g(x_k^{k+m-1})\hat{P}_n(x_0^{m-1}, x_k^{k+m-1}),$$

where $P(x_0^{m-1}, x_k^{k+m-1}) = \mathbb{P}_P[X_0^{m-1} = x_0^{m-1}, X_k^{k+m-1} = x_k^{k+m-1}]$ for any probability measure P on $\mathcal{X}^{\mathbb{Z}}$; and analogously for $\mathbb{E}[g(X_0^{m-1})g(X_k^{k+m-1})]$. Thus, due to the boundedness of $g(.)$ and the finiteness of $\mathcal{X}^m$,

$$\begin{aligned}
&\max_{0\le k\le b_n} |\hat{R}_g(k) - R_g(k)| \\
\le\ & const. \max_{0\le k\le b_n; x_0^{m-1}, x_k^{k+m-1}\in\mathcal{X}^m} |\hat{P}_n(x_0^{m-1}, x_k^{k+m-1}) - P(x_0^{m-1}, x_k^{k+m-1})|. \quad \text{(B.14)}
\end{aligned}$$

As in the proof of Proposition B.2, we can restrict ourselves to a version $\tilde{P}_n$ of $\hat{P}_n$, which has context function $c_n(.)$ instead of $\hat{c}_n(.)$. Then,

$$P(x_0^m, x_k^{k+m-1}) = \sum_{x_m^{k-1}\in\mathcal{X}^{k-m}} \prod_{t=v}^{k+m-1} P(x_t|c(x_0^{t-1}))P(x_0^{v-1}), \quad \text{(B.15)}$$

$$\tilde{P}(x_0^m, x_k^{k+m-1}) = \sum_{x_m^{k-1}\in\mathcal{X}^{k-m}} \prod_{t=v}^{k+m-1} \frac{N(x_t c_n(x_0^{t-1}))}{N(c_n(x_0^{t-1}))} \frac{N(x_0^{v-1})}{n}, \quad \text{(B.16)}$$

where $v \le k+m$ is the smallest integer such that x_0^{v-1} is in τ_{c_n} [v depends on x_0^{k+m-1}].

We now follow the strategy from Step2 in the proof of Proposition B.1. In formula (B.15), the context function $c(.)$ can be replaced with the sieve approximation context function $c_n(.)$ plus an error term which is bounded by

$$\sum_{j=1}^{b_n+m-1} (\nu(p_n)(b_n+m-1)\mathrm{card}(\mathcal{X}|)^j = O(b_n n^{-1/2}) = O(\log(n)^2 n^{-1/2}), \quad \text{(B.17)}$$

due to the decay of $\nu(.)$ in (B3) and making the choice of $d_n = C_d \log(n)$ with $C_d \ge 1/(2\log(1/\xi))$ [either $p_n = d_n$ or $\nu(p_n) \equiv 0$ for all n sufficiently large]. The estimation error in the sieve approximation is governed by the quantity $\Delta = \sup_{x\in\mathcal{X}, w\in\tau_{c_n}} \max\{|N(xw)/N(w) - P(x|w)|, |n^{-1}N(w) - P(w)|\}$ which can be bounded as [see proof of Proposition B.2, formula B.8]

$$\Delta = O_P(p_{min}^{-1} \log(n)^{1/2} \log(\mathrm{card}(\tau_{c_n}))^{1/2} n^{-1/2}) = o_P(\log(n)^2 n^{-1/2}), \quad \text{(B.18)}$$

by the assumptions on $p_{min,T} \le p_{min}$ and $\mathrm{card}(\tau_{c_n})$. Again following the argument in Step2 in the proof of Proposition B.1 and using the bounds in (B.17) and (B.18) leads to

$$\begin{aligned}
&\max_{0\le k\le b_n; x_0^{m-1}, x_k^{k+m-1}\in\mathcal{X}^m} |\hat{P}_n(x_0^{m-1}, x_k^{k+m-1}) - P(x_0^{m-1}, x_k^{k+m-1})| \\
\le\ & \sum_{j=1}^{b_n+m-1} (\Delta(b_n+m-1)\mathrm{card}(\mathcal{X}))^j + O(\log(n)^2 n^{-1/2}) \\
=\ & O_P(b_n \log(n)^2 n^{-1/2}) = O_P(\log(n)^4 n^{-1/2}).
\end{aligned}$$

Together with (B.12)-(B.14) we complete the proof. □

Acknowledgments: Comments by an associate editor and both referees helped to clarify the statements and expositions of the results.

References

[1] Bird, A. (1987). CpG islands as gene markers in the vertebrate nucleus. Trends in Genetics **3**, 342–347.

[2] Bosq, D. (1996). Nonparametric Statistics for Stochastic Processes. Lect. Notes in Statist. **110**. Springer.

[3] Braun, J.V. and Müller, H.-G. (1998). Statistical methods for DNA sequence. Statist. Sci. **13**, 142–162.

[4] Bühlmann, P. (1996). Locally adaptive lag-window spectral estimation. J. Time Series Anal. **17**, 247–270.

[5] Bühlmann, P. and Künsch, H.R. (1999). Block length selection in the bootstrap for time series. Comput. Statist. Data Anal. **31**, 295–310.

[6] Bühlmann, P. and Wyner, A.J. (1999). Variable length Markov chains. Ann. Statist. **27**, 480–513.

[7] Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. Bull. Math. Bio. **51**, 79–94.

[8] Datta, S. and McCormick W.P. (1995). Some continuous Edgeworth expansions for Markov chains with applications to bootstrap. J. Multiv. Anal. **52**, 83–106.

[9] Doukhan, P. (1994). Mixing. Properties and Examples. Lect. Notes in Statist. **85**. Springer.

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

[11] Fahrmeir, L. and Tutz, G. (1994). Multivariate Statistical Modelling Based on Generalized Linear Models. Springer.

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

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

[14] Hall, P. (1987). On the bootstrap and continuity correction. J. Roy. Statist. Soc. B **49**, 82–89.

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

[16] Kulperger, R.J. and Prakasa Rao, B.L.S. (1989). Bootstrapping a finite state Markov chain. Sankhya Ser. A **51**, 178–191.

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

[18] Leung, M.-Y., Marsh, G.M. and Speed, T.P. (1996). Over- and underrepresentation of short DNA words in herpesvirus genomes. J. Comp. Bio. **3**, 345–360.

[19] Muri, F. (1998). Modelling bacterial genomes using hidden Markov models. In COMPSTAT'98, Proceedings in Computational Statistics (Eds. R. Payne and P. Green), pp.89–100. Physica-Verlag.

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

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

[22] Schbath, S. (1997). An efficient statistic to detect over- and under-represented words in DNA sequences. J. Comp. Bio **4**, 189–192.

[23] Schbath, S., Prum, B. and deTurckheim, E. (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. J. Comp. Bio. **2**, 417–437.

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

[25] Van der Vaart, A.W. and Wellner, J.A. (1996). Weak Convergence and Empirical Processes. Springer.

[26] Woodroofe, M. and Jhun, M. (1989). Singh's theorem in the lattice case. Statist. & Probab. Lett. **7**, 201–205.