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

Working Paper**Author(s):**

Bühlmann, Peter Lukas

Publication date:

1999

Permanent link:

<https://doi.org/10.3929/ethz-a-004159940>

Rights / license:

In Copyright - Non-Commercial Use Permitted

Originally published in:

Research report / Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) 89

SIEVE BOOTSTRAP WITH VARIABLE LENGTH MARKOV
CHAINS FOR STATIONARY CATEGORICAL TIME SERIES

by

PETER BÜHLMANN

Research Report No. 89
October 1999

Seminar für Statistik
Eidgenössische Technische Hochschule (ETH)
CH-8092 Zürich
Switzerland

SIEVE BOOTSTRAP WITH VARIABLE LENGTH MARKOV CHAINS FOR STATIONARY CATEGORICAL TIME SERIES

PETER BÜHLMANN

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

October 1999

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 but still parsimoniously parameterized class of Markov models. Then, the resampling is given by simulating from the fitted model.

We show that in the ‘niche’ of categorical time series, the VLMC-sieve outperforms the more general block bootstrap. Moreover, the VLMC-sieve scheme enjoys the implementational advantage of using the plug-in rule for bootstrapping a statistical procedure, which is generally not the case for the block method.

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.

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. 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 with a parsimonious parameterization. For more details about VLMC's, see Bühlmann and Wyner (1999), where also bootstrapping in a VLMC has been discussed. Here, the main new issues we discuss are:

- (i) the VLMC-based bootstrap can be viewed as a sieve bootstrap for *general* categorical time series,
- (ii) the VLMC-sieve bootstrap is *more accurate* than the block bootstrap.

There seem 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 to 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 a clear answer to this gap. 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 'CpG-islands' which are indicators for 'start' regions of genes [inhomogeneity regions] in a long DNA sequence, cf. Bird (1987).

An important implementational advantage of the VLMC-sieve bootstrap is given through its use of the practically very convenient plug-in rule: the same statistical procedure/program for the original sample is applied to the bootstrap resample. This is in contrast to the block bootstrap which generally doesn't use the plug-in rule and the computation of the block bootstrapped estimator may have to be redesigned.

Implementation and the new theoretical and empirical results in this paper are strong reasons for attractiveness of the VLMC-sieve bootstrap for categorical time series.

2 The VLMC-sieve bootstrap for stationary categorical time series

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

$$\begin{aligned} \hat{\theta}_n &= T(\hat{P}_{n,emp}^{(m)}), \\ \hat{P}_{n,emp}^{(m)}(x) &= (n - m + 1)^{-1} \sum_{t=1}^{n-m+1} 1_{[(X_t, \dots, X_{t+m-1})=x]} \quad (x \in \mathcal{X}^m, \quad m \in \mathbb{N}) \end{aligned} \quad (2.1)$$

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

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

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

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

where $g : \mathcal{X}^m \rightarrow \mathbb{R}^q$ ($q \geq 1$) is arbitrary and $f : \mathbb{R}^q \rightarrow \mathbb{R}$ continuously differentiable at $\mu_g = \mathbb{E}[g(X_t, \dots, X_{t+m-1})]$. 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 3.3]. 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. We denote by $x_i^j = x_j, x_{j-1}, \dots, x_i$ ($i < j$, $i, j \in \mathbb{Z} \cup \{-\infty, \infty\}$) a string whose components are written in reverse order, $wu = (w_{|w|}, \dots, w_2, w_1, u_{|u|}, \dots, u_2, u_1)$ is the concatenation of the strings w and u . 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 $c : \mathcal{X}^\infty \rightarrow \cup_{j=0}^\infty \mathcal{X}^j$ ($\mathcal{X}^0 = \emptyset$) a (variable projection) function which maps

$$\begin{aligned} c : x_{-\infty}^0 &\mapsto x_{-\ell+1}^0, \quad \text{where } \ell \text{ is defined by} \\ \ell &= \ell(x_{-\infty}^0) = \min\{k; \mathbb{P}[X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0] = \mathbb{P}[X_1 = x_1 | X_{-k+1}^0 = x_{-k+1}^0] \text{ for all } x_1 \in \mathcal{X}\}, \\ &\text{where } \ell \equiv 0 \text{ corresponds to independence.} \end{aligned}$$

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

$$|c(x_{-\infty}^0)| = \ell(x_{-\infty}^0) \leq p \text{ for all } x_{-\infty}^0 \in \mathcal{X}^\infty.$$

The number p is called the order of the context function $c(\cdot)$, 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 $0, -1, \dots$ 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 $c(\cdot)$. Clearly, a VLMC of order p is a Markov chain of order p , now having a *memory of variable length* ℓ . If the context function $c(\cdot)$ 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.

Definition 2.2 Let $c(\cdot)$ be a context function of a stationary VLMC. The $(|\mathcal{X}|$ -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 $c(\cdot)$ can be reconstructed from τ_c . The context tree τ_c , which does not have to be complete with $|\mathcal{X}|$ 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

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

rectangle symbolizing the absent nodes 0,1 and 2 in depth 2, which can be thought as a completion of the tree with nodes lumped together to the state 3.

Fitting a VLMC involves a version of the tree structured context algorithm [Rissanen, 1983] for estimating the variable length memory, described by the context function $c(\cdot)$, 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 $\hat{\tau}_c$ is available, transition probabilities are determined by relative frequencies. The exact description of the algorithm as used here can be found in Bühlmann and Wyner (1999). The context algorithm yields a consistent estimate

$$\hat{P}_n = \hat{P}_{n;VLMC} \quad (2.3)$$

for the distribution 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 core for these results is consistency of the context algorithm for finding the true underlying structure of the context function $c(\cdot)$ from a general stationary process P . Generally $c(\cdot)$ is of infinite order. We consider then the truncated context function

$$c_n(x_{-\infty}^0) = \begin{cases} c(x_{-\infty}^0) & \text{if } w = c(x_{-\infty}^0) \text{ has length } |w| \leq d_n \\ x_{-d_n+1}^0 & \text{if } |c(x_{-\infty}^0)| > d_n \end{cases}, \quad (2.4)$$

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

$$\hat{c}_n \text{ from the context algorithm,} \quad (2.5)$$

satisfies

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

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

2.2 Definition of the VLMC-sieve bootstrap

The construction of the VLMC-sieve bootstrap is now straightforward by simulating from $\hat{P}_n$ in (2.3),

$$X_1^*, \dots, X_n^* \sim \hat{P}_n.$$

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

$$\hat{\theta}_n^* = h_n(X_1^*, \dots, X_n^*), \quad (2.6)$$

with h_n the defining function of the estimator $\hat{\theta}_n = h_n(X_1, \dots, X_n)$. The fact that the VLMC-sieve bootstrap employs the plug-in rule is a substantial 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 $\hat{\theta}_n^*$. For the estimator in (2.1), the following representation then holds,

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

2.3 Facts about the VLMC-sieve bootstrap

We describe here informally theoretical properties of the VLMC-sieve bootstrap, partially explaining theoretical superiority over the block bootstrap [Künsch, 1989]. Rigorous details are given in section 4.

Fact 2.1 *Assume that the data generating process $(X_t)_{t \in \mathbb{Z}} \sim P$ is stationary and suitably mixing [i.e., asymptotic independence of values as separation lag increases]. Then, the VLMC-sieve bootstrap is consistent for estimators as defined in (2.1) with T smooth at $P^{(m)}$.*

See Theorem 4.1. Fact 2.1 guarantees the basic requirement of consistency. In the following we describe some accuracy results. We restrict the focus to estimators

$$\hat{\theta}_n = (n - m + 1)^{-1} \sum_{t=1}^{n-m+1} g(X_t, \dots, X_{t+m-1}), \quad (m \in \mathbb{N}) \quad (2.7)$$

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

Fact 2.2 *Assume that the data generating process is as in Fact 2.1 with exponentially decaying mixing coefficients [i.e. exponentially fast decreasing dependence between values as separation lag increases], and satisfies additionally a sparseness condition for its potentially infinitely long memory. Then, for $\hat{\theta}_n$ as in (2.7),*

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

See Theorem 4.2. The value of Fact 2.2 is a rate-‘adaptivity’ property 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,

$$\mathbb{E} \left[\left(n \text{Var}^{B*}(\hat{\theta}_n^{B*}) - n \text{Var}(\hat{\theta}_n) \right)^2 \right] \sim \text{const.} n^{-2/3}, \quad (2.8)$$

achieved with the optimal blocklength parameter $\ell = \text{const.} n^{1/3}$. 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. In this sense, the blockwise bootstrap is not adaptive: its rate for variance estimation 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 $q : \mathbb{R} \rightarrow \mathcal{X}$, $X_t = q(Y_t)$, where $(Y_t)_{t \in \mathbb{Z}}$ is an exponentially mixing $\mathbb{R}$ -valued sequence.

The quality of the bootstrap approximation for the distribution of $\hat{\theta}_n$ hinges also on the ability to estimate the true underlying skewness.

Fact 2.3 *Under the assumptions of Fact 2.2, the VLMC-sieve bootstrap estimates the skewness of $\hat{\theta}_n$ very accurately. The rate of convergence [for the suitably standardized skewness] is bounded by $O_P(\log(n)^8 n^{-1/2})$.*

See Theorem 4.3. As in Fact 2.2, we achieve the finite parametric bound up to the $\log(n)^8$ term which is substantially better than for the block bootstrap.

3 Numerical examples

We analyze whether the Facts from section 2.3 in asymptotic theory can be explored for finite samples.

3.1 Variance estimation in simulations

The differences in variance estimation between Fact 2.2 [or Theorem 4.2] and (2.8) can be well seen in finite-sample problems, and the gain of the VLMC-sieve bootstrap is often substantial.

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

$$X_t = 1_{[Y_t > 0]}, \quad Y_t = 0.8Y_{t-1}1_{[Y_{t-1} > 0]} + \varepsilon_t, \quad (t \in \mathbb{Z}), \quad (3.1)$$

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],

$$X_t = 1_{[Y_t > 0]}, \quad Y_t = 0.8Y_{t-1} + \varepsilon_t, \quad (t \in \mathbb{Z}), \quad (3.2)$$

where $(\varepsilon_t)_{t \in \mathbb{Z}}$ is an i.i.d. innovation sequence, $\varepsilon_t \sim t_6$ independent from $\{Y_s; s < t\}$. Of interest are the stationary binary processes $(X_t)_{t \in \mathbb{Z}}$ in (3.1) and (3.2). The sample size is $n = 512$. As a simple estimator, we consider $\hat{\theta} = \bar{X}_n = n^{-1} \sum_{t=1}^n X_t$.

Figure 3.1: Bootstrap variance estimation of $n^{1/2}\bar{X}_n$ by $n \text{Var}^*(\bar{X}_n^*)$. Sample size $n = 512$. Left panel: model (3.1); right panel: model (3.2). The target $n \text{Var}(\bar{X}_n)$ is indicated by the horizontal lines. VLMC-sieve bootstrap indicated by ‘S’ with cutoffs as $\chi^2_{1;\alpha}/2$ with $\alpha = 0.95, 0.97, 0.99$; block bootstrap indicated by ‘B’ with $\ell = 8$ and estimated $\hat{\ell}$. 100 simulation runs, 500 bootstrap replicates per simulation run.

The VLMC-sieve bootstrap is run with different cutoff tuning parameters in the context algorithm for constructing the estimate $\hat{P}_n = \hat{P}_{n;VLMC}$. This cutoff parameter governs a tree pruning procedure and thus a model selection procedure within the class of VLMC’s. 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 (2.8)], 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 3.1 and Table 3.1. Overall, the VLMC-sieve bootstrap with 95% or 97% cutoff is significantly better than any of the two block bootstraps. In the sparse model (3.1), 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 (3.2), the best VLMC-sieve bootstrap is with 95% cutoff. It is better, with a high degree of significance, than the better of the two block methods, i.e. using the estimated $\hat{\ell}$. The block bootstrap is 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 $n = 128$. For example, the 95% cutoff VLMC is clearly better than the block bootstrap with estimated $\hat{\ell}$, with

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

Table 3.1: P -values from paired Wilcoxon-tests for MSE of bootstrap variance estimates $n \text{Var}^*(\bar{X}_n^*)$. On top: model (3.1); at bottom: model (3.2). Sample size $n = 512$. VLMC-sieve bootstrap indicated by ‘S’ with cutoffs as $\chi_{1;\alpha}^2/2$ with $\alpha = 0.95, 0.97, 0.99$; block bootstrap indicated by ‘B’ 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.

resulting P -value [as in Table 3.1] equal to 0.007 in model (3.1) and 0.002 in model (3.2), respectively.

3.2 Coverage of confidence intervals in simulations

We also examine construction of two-sided confidence interval with the estimator $\hat{\theta} = \bar{X}_n$ for $\theta = \mathbb{E}[X_t] = 0.725$ in model (3.1) and $\theta = 0.5$ in model (3.2) on nominal coverage level 0.9. The sample size is now chosen as $n = 128$.

We consider first order VLMC and block bootstraps and corrections thereof. Due to the discreteness of observations X_t [latticeness of the problem], correction of confidence intervals from an asymptotic point of view might not yield much or even no improvement [although with time series, the variance estimate of $n \text{Var}(\hat{\theta})$, which determines the accuracy of first order intervals, is typically less accurate than the typical order $1/\sqrt{n}$ of error due to latticeness]. However, there might be still some [considerable] gain to employ corrections for finite samples. For a related discussion see Hall (1987) and Woodroffe and Jhun (1989). For the VLMC method we examine the double bootstrap. Consider construction of a two-sided confidence interval which covers θ with 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

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

where $\theta^* = T(\hat{P}_n^{(m)})$ with T and $\hat{P}_n$ as in (2.1) and (2.3), respectively. Now consider an additive correction of the original nominal coverage level by using the double bootstrap. Based on $X_1^*, \dots, X_n^*$, run the VLMC-sieve bootstrap to obtain $X_1^{**}, \dots, X_n^{**} \sim \hat{P}_n^*$. Now,

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

Sparse	VLMC, 95%	Block, $\hat{\ell}$	Double VLMC, 95%	Block BC _a , $\hat{\ell}$
coverage	0.81	0.75	0.88	0.78
median(length)	0.178	0.164	0.215	0.164
MAD(length)	0.034	0.023	0.065	0.022
Non-sparse	VLMC, 95%	Block, $\hat{\ell}$	Double VLMC, 95%	Block BC _a , $\hat{\ell}$
coverage	0.74	0.71	0.84	0.75
median(length)	0.276	0.262	0.368	0.258
MAD(length)	0.041	0.063	0.041	0.122

Table 3.2: 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 (3.1); at bottom: model (3.2). First order accurate VLMC-sieve and block bootstrap intervals; and corrections with the double VLMC-sieve and the BC_a block bootstrap. Based on 100 simulations, 500 first level bootstrap replicates; VLMC double bootstrap calibration with 100 first and 100 second level bootstrap replicates.

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 $X_1, \dots, X_n$]. 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 $1 - \alpha$. Importantly, studentization which can be very inefficient in practice, isn't necessary. 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 3.2. For the block bootstrap, the BC_a method increases performance a bit compared to the first order block intervals, although the difference is not significant in the sparse model (3.1) and only weakly significant in the non-sparse model (3.2). The first order VLMC shows significant gain over the first order block bootstrap in model (3.1), whereas there is no significant difference in model (3.2). The double VLMC bootstrap technique has a significant advantage over all other methods in both models. On average, it corrects the nominal 90% to the 95.4% and 97.3% coverage level in models (3.1) and (3.2), respectively.

3.3 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

G-islands [Bird, 1987]. The chance for occurrence of the pair

of consecutive base elements (C, G) might not be constant over the whole DNA string and stretches with high chances are classified as interesting potential ‘start’ 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 (so-called junk DNA) which are ignored by the cell chemistry. Physically, the genes are spaced apart and separated by junk DNA which we term ‘inter-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,

$$\begin{aligned}\hat{\theta}_n(t) &= \sum_{j=t-499}^{t+499} w_j(t) 1_{[X_j=C, X_{j+1}=G]}, \\ w_j(t) &= h_j(t) / \sum_{j=t-499}^{t+499} h_j(t), \quad h_j(t) = \max\{1 - |t - j|/500, 0\} \quad (t = 500, \dots, n - 500)\end{aligned}$$

is shown in Figure 3.2. An alternative is the z-score statistic,

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

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 $\hat{P}_n$, assuming that $\hat{P}_n$ is stationary]. Such z-scores but 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, $\text{Var}(\hat{\theta}_n(t) - \theta^*) = \text{Var}(\hat{\theta}_n(t))(1 + o(1))$ 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 smoothed frequencies $\hat{\theta}_n(t)$ or z-scores $\hat{\zeta}_n(t)$, the choice of the non-flexible full Markov model is 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 $\text{Var}^*(\hat{\theta}_n^*(t))$ [and θ^*] might be extremely poor. Our VLMC 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 that $\theta(t)$ isn’t constant and thus P not stationary. Under the hypothesis, the nuisance parameter P is consistently estimated by an approximating VLMC, denoted by $\hat{P}_n$. Then, the distribution under the hypothesis of $\hat{\theta}_n(t)$ [which is the same for all t] is estimated by the VLMC-sieve bootstrap $\hat{\theta}_n^*(500)$ and acceptance regions are then determined from the bootstrap quantiles. Similarly, 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.

Figure 3.2: Smoothed frequencies $\hat{\theta}_n(t)$ [top] and z-scores $\hat{\zeta}_n(t)$ [bottom] of pair *CG* indicated by solid line. One- and two-sided acceptance regions on 5% significance level indicated by coarse and fine dashed lines, respectively [one-sided against alternative of large frequencies or z-scores].

One- and two-sided acceptance regions on the 5% significance level are displayed in Figure 3.2. Using smoothed frequencies $\hat{\theta}_n(t)$ and its bootstrap quantiles or z-scores $\hat{\zeta}_n(t)$ with its standard normal quantiles give similar results. The bootstrap distribution of $\hat{\theta}_n^*(500)$ is close to normal and we conclude that the bootstrap variance must be close to the true variance; although the corresponding window size is only moderate with 1000 observations. Potential locations of *CpG* 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; different periods of sticky character and switches can then be still picked up by the VLMC model.

4 Theoretical properties

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 $P(x|w) = P(xw)/P(w)$ ($x, w \in \cup_{j=1}^{\infty} \mathcal{X}^j$).

4.1 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

$$\begin{aligned} P_n \text{ is a VLMC with context function } c_n(\cdot), \\ \text{and with transition probabilities } P_n(x|w) = \begin{cases} P(xw)/P(w) & \text{if } P(w) \neq 0 \\ 0 & \text{otherwise} \end{cases}, \\ \text{for } w \in \tau_{c_n}, \text{ where } c_n(\cdot) \text{ is defined in (2.4).} \end{aligned} \quad (4.1)$$

We denote by p_n the order of $c_n(\cdot)$ 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 $c_n(\cdot)$], and strictly smaller than d_n for all n sufficiently large only if P is Markovian of finite order. As we will assume in (A2) and (A3) below, $d_n = \text{const.} \log(n)$ 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

$$\nu(k) = \nu_P(k) = \sup_{x_{-\infty}^1 \in \mathcal{X}^\infty} |P(x_1|x_{-k+1}^0) - P(x_1|x_{-\infty}^0)| \quad (k \in \mathbb{N}). \quad (4.2)$$

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 (4.2) is the right notion for describing the quality of a Markov-sieve approximation.

Remark 4.1. 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 $c_n(\cdot)$. Also, $\nu_P(k) = 0$ for all $k \geq p$.

Estimation of P_n is based on the context algorithm

$$\begin{aligned} \hat{P}_n \text{ with context function } \hat{c}_n(\cdot), \\ \text{and transition probabilities } \hat{P}_n(x|w) = N(xw)/N(w) \text{ for all } w \in \tau_{\hat{c}_n}. \end{aligned} \quad (4.3)$$

Here and in the sequel $N(w) = \sum_{t=|w|}^n 1_{[X_{t-|w|+1}^t=w]}$.

The following assumptions guarantee good behavior of $\hat{P}_n$ in (4.3) and in turn also for the VLMC-sieve bootstrap in (2.6). The context tree τ_{c_n} induced by $c_n(\cdot)$ [Definition 2.2] includes the terminal node context tree

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

(A1) The process $(X_t)_{t \in \mathbb{Z}} \sim P$ is stationary and geometrically α -mixing with $\alpha_P(k) \leq \text{const.} \rho^k$ for some $0 \leq \rho < 1$.

(A2) The sieve approximations P_n as defined in (4.1), having context functions $c_n(\cdot)$ with truncation d_n as defined in (2.4), satisfy: For any $d_n \sim C_d \log(n)$ with $0 < C_d < \infty$,

$$|\tau_{c_n}| = o(n),$$

$p_{min,T}^{-1} = O(n^\beta)$ for some $0 \leq \beta < 1/2$, where $p_{min,T} = \min_{w \in \tau_{c_n}^T} P(w)$,

$p_{sep,T}^{-2} = O\left(\frac{(np_{min,T}^{(1-\gamma)/2})^{1-\gamma}}{\log(n)^{1+\delta}}\right)$ for some $0 \leq \gamma < 1$, $\delta > 0$,

where $p_{sep,T} = \min_{wu \in \tau_{c_n}^T, u \in \mathcal{X}} \sum_{x \in \mathcal{X}} |P(x|wu) - P(x|w)|$,

$\left(\min_{x \in \mathcal{X}, w \in \tau_{c_n}^T} P(x|w)\right)^{-1} = O(n)$.

(A3) For any $d_n \sim C_d \log(n)$ 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 $c_n(\cdot)$.

(A4) $\nu_P(k) \leq \text{const.} k^{-1+\iota}$ ($\iota > 0$) for all $k \in \mathbb{N}$.

Assumption (A1) and (A2) ensure consistent estimation of $c_n(\cdot)$ with the context algorithm, see Ferrari (1999); (A3) is a weak condition since λ can be much larger than 1; (A4) is a mild requirement on the decay of the dependence.

4.2 Consistency and accuracy of the VLMC-sieve bootstrap

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 4.1 *Let $(X_t)_{t \in \mathbb{Z}}$ be a stationary $\mathcal{X}$ -valued process satisfying (A1)-(A4). Let $\hat{\theta}_n$ be as in (2.1) with T Hadamard differentiable at $P^{(m)}$ with continuous derivative, let $\hat{\theta}_n^*$ be its sieve-bootstrapped version defined in (2.6) and assume that [the existing limiting variance] $\lim_{n \rightarrow \infty} n \text{Var}(\hat{\theta}_n)$ 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 $\theta^* = T(\hat{P}_n^{(m)})$.

A proof is given in section 6. An estimator which satisfies the Hadamard differentiability assumption is given in (2.2).

We provide now the order of the accuracy of bootstrap variance and skewness estimates.

Theorem 4.2 *Let $(X_t)_{t \in \mathbb{Z}} \sim P$ be a $\mathcal{X}$ -valued process, satisfying (A1)-(A4). Moreover, we strengthen part of (A2) to $p_{min,T}^{-1} = O(p_n)$ for p_n as in (A3), and (A4) to $\nu_P(k) \leq \text{const.} \xi^k$ for some $0 \leq \xi < 1$. Let $\hat{\theta}_n$ be as in (2.7) and $\hat{\theta}_n^*$ its sieve-bootstrapped version from (2.6). Then,*

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

A proof is given in section 6. The condition $p_{min,T}^{-1} = O(p_n)$ can be interpreted as a sparseness condition implying that $|\tau_{c_n}^T| \leq p_{min,T}^{-1} = O(p_n)$ 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 (3.1) satisfies this condition.

The quality of the bootstrap approximation for the distribution of $\hat{\theta}_n$ hinges also on the ability to estimate the true underlying skewness.

Theorem 4.3 *Assume the conditions and notation from Theorem 4.2. Moreover, denote by $\mu_{3;n} = \mathbb{E}[(n^{1/2}(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]))^3]$ and $\mu_{3;n}^* = \mathbb{E}^*[(n^{1/2}(\hat{\theta}_n^* - \mathbb{E}^*[\hat{\theta}_n^*]))^3]$. Then*

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{1/2} \mu_{3;n} &= \mu_{3;as}, \\ n^{1/2} \mu_{3;n}^* - n^{1/2} \mu_{3;n} &= O_P(\log(n)^8 n^{-1/2}) \quad (n \rightarrow \infty), \end{aligned}$$

where $\mu_{3;as} = \sum_{s,t=-\infty}^{\infty} \mathbb{E}[Z_0 Z_s Z_t]$ with $Z_t = g(X_t, \dots, X_{t+m-1}) - \mathbb{E}[g(X_t, \dots, X_{t+m-1})]$.

A proof is given in section 6.

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. The sieve bootstrap is from a practical point of view 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: this is asymptotically described by Fact 2.2 [Theorem 4.2] and Fact 2.3 [Theorem 4.3], and it is also empirically found for finite samples. Particularly, the sieve bootstrap is ‘rate-adaptive’ with respect to the strength of dependence in the data generating process; this is practically relevant whenever only a ‘weak form’ of dependence is present.

Another advantage of the VLMC-sieve bootstrap is the construction of bootstrapped estimators with the plug-in rule, see formula (2.6): 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 $\hat{\theta}^{*B}$ may should be redesigned which can be an awkward task. On the other hand, the block resampling stage is very easy whereas the tree structured context algorithm for fitting a VLMC is quite sophisticated; however, the latter can be efficiently implemented with $O(n \log(n))$ essential operations.

6 Proofs

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

Lemma 6.1 Assume that $(X_t)_{t \in \mathbb{Z}} \sim P$ satisfies (A1). Let $\mathcal{A}$ be a set of finite sequences $w = x_{-|w|+1}^0$ and let $p_{\mathcal{A}} \geq \max\{|w|; w \in \mathcal{A}\}$ be an upper bound for the length of the elements in $\mathcal{A}$. $\mathcal{A}$ and $p_{\mathcal{A}}$ are allowed to depend on n . Assume that $p_{\mathcal{A}} = o(n)$ and $|\mathcal{A}| = o(n^r)$ for some $0 < r < \infty$. Then,

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

Proof: The bias is bounded as

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

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

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

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 $q = \min\{[C_q n / \log(n)], [C_q n / p_{\mathcal{A}}]\}$ for some suitable constant $0 < C_q < \infty$, we then obtain

$$\begin{aligned} & \mathbb{P}[a_n^{-1} |n^{-1}N(w) - \mathbb{E}[N(w)]| > \varepsilon] \\ & \leq \text{const.} |\mathcal{A}|^{-C_q \varepsilon^2 / 8} + \text{const.} \varepsilon^{-1/2} n^{5/4} \exp(\log(\rho) \log(n) / (4C_q)). \end{aligned} \quad (6.1)$$

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

$$|\mathcal{A}| n^{5/4} \exp(\log(\rho) \log(n) / (4C_q)) \rightarrow 0 \quad (n \rightarrow \infty), \quad (6.2)$$

due to the assumption $|\mathcal{A}|/n^r \rightarrow 0$ for some $0 < r < \infty$. Now observe that

$$|\mathcal{A}| |\mathcal{A}|^{-C_q \varepsilon^2 / 8} \rightarrow 0 \quad (\varepsilon \rightarrow \infty). \quad (6.3)$$

Therefore, by (6.1)-(6.3) we have

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

which completes the proof. $\square$

We now prove that the sieve approximation P_n from (4.1) and its estimated version from (4.3) satisfy a Doeblin-type condition. Let

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

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

Proposition 6.1 Assume that $(X_t)_{t \in \mathbb{Z}} \sim P$ satisfies (A1), (A3) and (A4). 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

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

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

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

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

By using (A1), (A3) and choosing C_r sufficiently large [note that $p_n \leq d_n = C_d \log(n) = O(\log(n))$], this implies that there exists an n_0 with

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

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,

$$P^{(r_n)}(A, x) = \sum_{x_{r_n-p_n+1}^{r_n} \in A} \sum_{x_{p_n}^{r_n-p_n} \in \mathcal{X}^{r_n-1}} P(x_{r_n} | x_0^{r_n-1}) P(x_{r_{n-1}} | x_0^{r_n-2}) \cdots P(x_{p_n} | x_0^{p_n-1}). \quad (6.4)$$

And the same expansion also holds for $P_n^{(r_n)}(A, x)$ in terms of conditional probabilities $P_n(x_t | x_0^{t-1})$ ($p_n \leq t \leq r_n$). 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 $P_n(x_t | x_0^{t-1})$ ($p_n \leq t \leq r_n$), 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) $|c^{(ext)}(x_0^{t-1})| > |c_n^{(ext)}(x_0^{t-1})|$ for all n sufficiently large; (III) $|c^{(ext)}(x_0^{t-1})| < |c_n^{(ext)}(x_0^{t-1})|$ for all n sufficiently large. Case (III) is impossible by definition of $c_n(\cdot)$. 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 $\sup_{t, x_0^t} |\Delta(x_0^t)| \leq \nu(p_n)$. Now using that $\sum_{x_t} P(x_t | x_0^{t-1}) = 1$ for all x_0^{t-1} , we obtain

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

which tends to zero by (A4) [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. $\square$

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

Proposition 6.2 *Assume that $(X_t)_{t \in \mathbb{Z}} \sim P$ satisfies (A1)-(A4). 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 $\mathbb{P}[B_n] \rightarrow 1$ such that*

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

Proof: Denote by $\hat{c}_n(\cdot)$ the estimated context function. The results in Ferrari (1999) show that there exists a sequence of sets D_n with $\mathbb{P}[D_n] \rightarrow 1$ such that $\hat{c}_n(\cdot) = c_n(\cdot)$ 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 $c_n(\cdot)$ instead of $\hat{c}_n(\cdot)$. 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

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

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 (A2). Applying Lemma 6.1 and using $|\tau_{c_n}| = o(n)$ from (A2) 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,

$$\Delta = o_P(p_{min}^{-1} \log(n)n^{-1/2}) = o_P(n^{\beta-1/2} \log(n)). \quad (6.5)$$

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 6.1: the bound becomes

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

since $\beta < 1/2$ from (A2) implies $\Delta r_n = o_P(n^{\beta-1/2} \log(n)^2) = o_P(1)$. Therefore, using the result from Proposition 6.1, we can find a sequence of sets B_n [which are subsets of D_n] with $\mathbb{P}[B_n] \rightarrow 1$ on which the first assertion holds. The second assertion is an immediate consequence of the first. $\square$

Proof of Theorem 4.1.

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

$$n^{1/2}(\hat{P}_{n,emp}^{(m)} - P^{(m)}) \Rightarrow W, \quad (6.6)$$

$$n^{1/2}(\hat{P}_{n,emp}^{*(m)} - P^{*(m)}) \Rightarrow W \text{ in probability,} \quad (6.7)$$

where ‘ $\Rightarrow$ ’ 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 (6.6) follows by (A1). For (6.7), we first show convergence of covariances

$$n \text{Cov}^*(\hat{P}_{n,emp}^{*(m)}(x), \hat{P}_{n,emp}^{*(m)}(y)) - n \text{Cov}(\hat{P}_{n,emp}^{(m)}(x), \hat{P}_{n,emp}^{(m)}(y)) = o_P(1), \quad x, y \in \mathcal{X}^m. \quad (6.8)$$

This can be shown using the geometric mixing property of $\hat{P}_n$ [given in Proposition 6.2], analogously to the proof of Theorem 4.2 [which gives a good convergence rate but needs

more assumptions on $p_{min,T}$]. Finally, weak convergence in (6.7) 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 6.2, is sufficient to treat them as negligible. $\square$

Proof of Theorem 4.2.

Choose $r_n = C_r \log(n)$ as in Proposition 6.2 with $1 < C_r < \infty$ sufficiently large. Then set $b_n = \lceil C_b \log(n) r_n \rceil$ with $C_b \geq -1/(2 \log(1 - \kappa))$. Under the geometric mixing condition for P [see assumption (A1)], the variance of $\hat{\theta}_n$ can be written as

$$n \operatorname{Var}(\hat{\theta}_n) = \sum_{k=-b_n}^{b_n} R_g(k) + O(n^{-1/2}), \quad (6.9)$$

where $R_g(k) = \operatorname{Cov}(g(X_0^{m-1}), g(X_k^{k+m-1}))$, cf. Doukhan (1994, Ch.1.2.2). Likewise, using the geometric mixing condition for the bootstrap $\hat{P}_n$ as stated in Proposition 6.2,

$$n \operatorname{Var}^*(\hat{\theta}_n^*) = \sum_{k=-b_n}^{b_n} \hat{R}_g(k) + O_P(\log(n) n^{-1/2}), \quad (6.10)$$

where $\hat{R}_g(k) = \operatorname{Cov}^*(g((X^*)_0^{m-1}), g((X^*)_k^{k+m-1}))$. Formula (6.10) can be seen to hold as follows: on the set B_n [as stated in Proposition 6.2],

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

Due to the choice of C_b and r_n , this yields the bound $O_P(n^{-1/2} \log(n))$. By formulae (6.9) and (6.10), it remains to analyze

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

Write

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

where the O_P -term is uniform over k , arising from the difference $\mathbf{E}^*[g(X^*)_0^m]^2 - \mathbf{E}[g(X_0^m)]^2$ [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,

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

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 $\mathbf{E}[g(X_0^{m-1})g(X_k^{k+m-1})]$. Thus, due to the boundedness of $g(\cdot)$ and the finiteness of $\mathcal{X}^m$,

$$\begin{aligned} & \max_{0 \leq k \leq b_n} |\hat{R}_g(k) - R_g(k)| \\ & \leq \text{const.} \max_{0 \leq k \leq b_n; x_0^{m-1}, x_k^{k+m-1} \in \mathcal{X}^m} |\hat{P}_n(x_0^{m-1}, x_k^{k+m-1}) - P(x_0^{m-1}, x_k^{k+m-1})|. \end{aligned} \quad (6.11)$$

As in the proof of Proposition 6.2, we can restrict ourselves to a version $\tilde{P}_n$ of $\hat{P}_n$, which has context function $c_n(\cdot)$ instead of $\hat{c}_n(\cdot)$. Then,

$$P(x_0^m, x_k^{k+m-1}) = \sum_{x_m^{k-1} \in \mathcal{X}^{k-m}} \prod_{t=v}^{k+m-1} P(x_t | c(x_0^{t-1})) P(x_0^{v-1}), \quad (6.12)$$

$$\tilde{P}(x_0^m, x_k^{k+m-1}) = \sum_{x_m^{k-1} \in \mathcal{X}^{k-m}} \prod_{t=v}^{k+m-1} \frac{N(x_t c_n(x_0^{t-1}))}{N(c_n(x_0^{t-1}))} \frac{N(x_0^{v-1})}{n}, \quad (6.13)$$

where $v \leq k+m$ 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 6.1. In formula (6.12), the context function $c(\cdot)$ can be replaced with the sieve approximation context function $c_n(\cdot)$ plus an error term which is bounded by

$$\sum_{j=1}^{b_n+m-1} (\nu(p_n)(b_n+m-1)|\mathcal{X}|)^j = O(b_n n^{-1/2}) = O(\log(n)^2 n^{-1/2}), \quad (6.14)$$

due to the decay of $\nu(\cdot)$ in (A4) and making the choice of $d_n = C_d \log(n)$ with $C_d \geq 1/(2 \log(1/\xi))$ [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 6.2, formula 6.5]

$$\Delta = O_P(p_{min}^{-1} \log(n)^{1/2} \log(|\tau_{c_n}|)^{1/2} n^{-1/2}) = o_P(\log(n)^2 n^{-1/2}), \quad (6.15)$$

by the assumptions on $p_{min,T} \leq p_{min}$ and $|\tau_{c_n}|$. Again following the argument in Step2 in the proof of Proposition 6.1 and using the bounds in (6.14) and (6.15) leads to

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

Together with (6.9)-(6.11) we complete the proof. $\square$

Proof of Theorem 4.3.

From covariance bounds for α -mixing sequences, cf. Doukhan (1994, Ch.1.2.2), we obtain

$$\mathbb{E}[Z_0 Z_s Z_t] \leq \max(\alpha_P(|s| - m + 1), \alpha_P(|t| - m + 1), \alpha_P(|s - t| - m + 1)) \quad (s, t \in \mathbb{Z}).$$

Due to the geometric α -mixing in property of P in (A1), this then implies the limit $\mu_{3;as}$ for $n^{1/2} \mu_{3;n}$. Furthermore, using also the geometric mixing property of $\hat{P}_n$ for the bootstrapped process $(X_t^*)_{t \in \mathbb{Z}}$ as stated in Proposition 6.2, we proceed analogously as in the proof of Theorem 4.2, formulae (6.9) and (6.10), so that for b_n as in the proof of Theorem 4.2, we only need to control

$$\max_{0 \leq s, t \leq b_n} |\mathbb{E}^*[Z_0^* Z_s^* Z_t^*] - \mathbb{E}[Z_0 Z_s Z_t]|.$$

But this can be bounded, due to the boundedness of $g(\cdot)$ and the finiteness of $\mathcal{X}$, by a constant times

$$\max_{0 \leq s, t \leq b_n; x_0^{m-1}, x_s^{s+m-1}, x_t^{t+m-1} \in \mathcal{X}^m} |\hat{P}_n(x_0^{m-1}, x_s^{s+m-1}, x_t^{t+m-1}) - P(x_0^{m-1}, x_s^{s+m-1}, x_t^{t+m-1})|.$$

Using the same argument as in the proof of Theorem 4.2, the latter can be bounded as $O_P(\log(n)^4 n^{-1/2})$. Summation in the range $0 \leq s, t \leq b_n = \text{const.} \log(n)^2$ [analogously to (6.9)] finally yields the additional $O(\log(n)^4)$ in the bound of the Theorem. $\square$

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. and Künsch, H.R. (1999). Block length selection in the bootstrap for time series. To appear in *Comput. Statist. Data Anal.*
- [5] Bühlmann, P. and Wyner, A.J. (1999). Variable length Markov chains. *Ann. Statist.* **27**, 480–513.
- [6] Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Bio.* **51**, 79–94.
- [7] Datta, S. and McCormick W.P. (1995). Some continuous Edgeworth expansions for Markov chains with applications to bootstrap. *J. Multiv. Anal.* **52**, 83–106.
- [8] Doukhan, P. (1994). *Mixing. Properties and Examples*. Lect. Notes in Statist. **85**. Springer.
- [9] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.
- [10] Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer.
- [11] Ferrari, F. (1999). Estimation of general stationary processes by variable length Markov chains. Preprint. ETH Zürich.
- [12] Götze, F. and Künsch, H.R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. *Ann. Statist.* **24**, 1914–1933.
- [13] Hall, P. (1987). On the bootstrap and continuity correction. *J. Roy. Statist. Soc. B* **49**, 82–89.

- [14] Kulperger, R.J. and Prakasa Rao, B.L.S. (1989). Bootstrapping a finite state Markov chain. *Sankhya Ser. A* **51**, 178–191.
- [15] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217–1241.
- [16] 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.
- [17] Muri, F. (1998). Modelling bacterial genomes using hidden Markov models. In *COMP-STAT'98, Proceedings in Computational Statistics* (Eds. R. Payne and P. Green), pp.89–100. Physica-Verlag.
- [18] Rissanen, J. (1983). A universal data compression system. *IEEE Trans. Inform. Theory* **IT-29**, 656–664.
- [19] Schbath, S. (1997). An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Com. Bio* **4**, 189–192.
- [20] 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.
- [21] Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer.
- [22] Woodroffe, M. and Jhun, M. (1989). Singh's theorem in the lattice case. *Statist. & Probab. Lett.* **7**, 201–205.