Model Selection for Variable Length Markov Chains and Tuning the Context Algorithm

by

Peter Bühlmann

Research Report No. 82
September 1997

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Model Selection for Variable Length Markov Chains and Tuning the Context Algorithm

Peter Bühlmann

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

September 1997

Abstract

We consider the model selection problem in the class of stationary variable length Markov chains (VLMC) on a finite space. The processes in this class are still Markovian of higher order, but with memory of variable length.

Various aims in selecting a VLMC can be formalized with different non-equivalent risks, such as final prediction error or expected Kullback-Leibler information. We consider the asymptotic behavior of different risk functions and show how they can be generally estimated with the same resampling strategy. Such estimated risks then yield new model selection rules: in the special case of classical higher order full Markov chains we obtain a better proposal than the AIC criterion, which has been suggested in the past.

Attacking the model selection problem also yields a proposal for tuning Rissanen's context algorithm, which can be used for estimating the minimal state space and in turn the whole probability structure of a VLMC.

Key words and phrases. Bootstrap, zero-one loss, final prediction error, Kullback-Leibler information, L_2 loss, optimal tree pruning, resampling.

Short title: Selecting variable length Markov chains

1 Introduction

We consider the model selection problem in the class of stationary variable length Markov chains (VLMC) on a finite space $\mathbf{X}$. The processes in this class are still Markovian of higher order, but their memory can have variable length. With a variable length memory, the minimal state space becomes smaller and unlike full high order Markov chains with fixed memory-length, the process is not heavily exposed to the curse of dimensionality. VLMC's are particularly attractive when there is long memory in certain 'directions'.

Estimation of the minimal state space and the probability distribution of a VLMC can be done with the tree structured context algorithm (Rissanen, 1983). This algorithm is consistent in very general situations, cf. Bühlmann and Wyner (1997). Moreover, it is known to be efficient in the sense of predictive coding, cf. Weinberger et al. (1995), and also in the statistical sense for estimating a smooth functional, cf. Bühlmann (1997). Successful applications of the context algorithm have been reported among others by Rissanen (1994) for modeling chaotic processes and by Weinberger et al. (1996) for data compression.

The model selection problem in the class of VLMC's is not well understood. Even for the classic full Markov chains of higher order the model selection problem has not been considered in more rigorous details. For estimation of the order of a full Markov chain, Tong (1975) has proposed the AIC criterion which should aim to minimize an expected Kullback-Leibler information. This proposal can be improved: our strategy of selecting a VLMC model is also a better proposal in the special case of order selection in classical full Markov chains. We study here the selection of a VLMC under different risk functions, such as final prediction error with the quadratic and the zero-one loss and the expected Kullback-Leibler information. The risks are not equivalent, by specifying a certain risk function we can tailor the model selection problem towards specific aims. The estimation of the various risks can be done consistently with a resampling scheme. As mentioned above in connection with order selection for full Markov chains, our method is not equivalent to the AIC criterion, even when using the expected Kullback-Leibler information as risk function.

To use the context algorithm mentioned above for fitting VLMC's one needs to choose a tuning parameter, the so-called cut-off. So far, this problem of tuning has not received any systematic attention. We discuss the relation of choosing the cut-off to model selection and so-called optimal tree pruning. Similar to selecting a model, we propose a resampling technique for estimating an optimal cut-off. The optimality of the cut-off is with respect to a chosen risk function, as in the model selection problem.

In section 2 we define the VLMC's and describe the context algorithm, in section 3 we show the behavior of different risks as a function of different estimated VLMC's, in section 4 we show how estimation of these risks can be done via resampling and discuss the tuning of the cut-off parameter for the context algorithm, in section 5 we present results from a simulation study, section 6 outlines some conclusions and in section 7 we give the proofs.

2 Variable length Markov Chains

In the sequel, we denote by $x_i^j = x_j, x_{j-1}, \ldots, x_i$ $(i < j,\ i, j \in \mathbb{Z} \cup \{-\infty, \infty\})$ a string written in reverse 'time'. We usually denote by capital letters X random variables and

by small letters x fixed deterministic values. We follow here the ideas of Weinberger et al. (1995) and define what we call a variable length Markov chain (VLMC). As a starting point, consider $(X_t)_{t\in\mathbb{Z}}$, being a stationary Markov chain of finite order k with values in a finite space $\mathbf{X}$. Thus,

$$\mathbb{P}[X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0] = \mathbb{P}[X_1 = x_1 | X_{-k+1}^0 = x_{-k+1}^0], \text{ for all } x_{-\infty}^1. \tag{2.1}$$

Such full Markov chains are very hard to estimate since they involve $|\mathbf{X}|^k(|\mathbf{X}|-1)$ free parameters. To get less complex models, the idea is to lump irrelevant states in the history X_{-k+1}^0 in formula (2.1) together, resulting in a sparse Markov chain.

For a time point $t \in \mathbb{Z}$, maybe only some values from the infinite history $X_{-\infty}^{t-1}$ of the variable X_t are relevant. This relevant history can be thought as a *context* for the actual variable X_t. To achieve a flexible model class, ranging from some type of sparse to full Markov chains, we let the length of a context depend on the actual values $X_{-\infty}^{t-1}$. For example, we might have for the variable X_t a context of length 1 and for $X_{t'}$ $(t' \neq t)$ a context of length 5. We can formalize this as follows.

Definition 2.1 *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process with values $X_t \in \mathbf{X}$, $|\mathbf{X}| < \infty$. Denote by $c : \mathbf{X}^\infty \to \mathbf{X}^\infty$ a (variable projection) function which maps*

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

Then, $c(.)$ 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.

The name *context* refers to the portion of the past that influences the next outcome. By the projection structure of the context function $c(.)$, the context-length $\ell(.) = |c(.)|$ determines $c(.)$ and vice-versa. The definition of ℓ implicitly reflects the fact that the context-length of a variable x_t is $\ell = |c(x_{-\infty}^{t-1})| = \ell(x_{-\infty}^{t-1})$, depending on the history $x_{-\infty}^{t-1}$.

Definition 2.2 *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process with values $X_t \in \mathbf{X}$, $|\mathbf{X}| < \infty$ and corresponding context function $c(.)$ as given in Definition 2.1. Let $0 \leq k \leq \infty$ be the smallest integer such that*

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

Then $c(.)$ is called a context function of order k, and $(X_t)_{t\in\mathbb{Z}}$ is called a stationary variable length Markov chain (VLMC) of order k. We always identify $(X_t)_{t\in\mathbb{Z}}$ with its probability distribution P_c on $\mathbf{X}^\infty$.

Clearly, a VLMC of order k is a Markov chain of order k, now having a *memory of variable length* ℓ. By requiring stationarity, a VLMC is thus completely specified by its transition probabilities,

$$p(x_1 | c(x_{-\infty}^0)) = \mathbb{P}_{P_c}[X_1 = x_1 | c(X_{-\infty}^0) = c(x_{-\infty}^0)], \; x_{-\infty}^1 \in \mathbf{X}^\infty.$$

In retrospect, we could define a context function $c(.) : \mathbf{X}^k \to \mathbf{X}^k$, since there is no functional dependence of the function $c(x_{-\infty}^0)$ on a variable x_{-k+1-m} $(m > 0)$. We sometimes

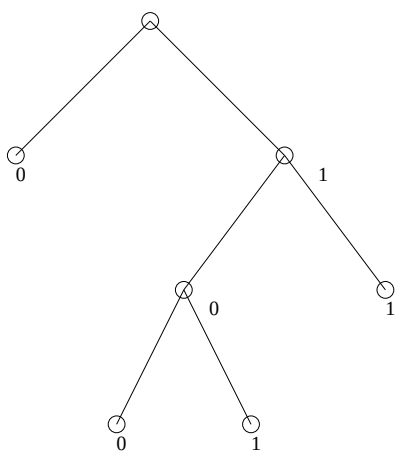
use the definition on $\mathbf{X}^\infty$ and sometimes on $\mathbf{X}^k$. The context function projects the k-th (or infinite) order history x_{-k+1}^0 into $\mathbf{X}^k$. Often the range space of the context function $c(.)$ is not the full space $\mathbf{X}^k$, but also not the empty space. If the context function $c(.)$ of order k is the full projection $x_{-k+1}^0 \mapsto x_{-k+1}^0$ for all x_{-k+1}^0, the VLMC is a full Markov chain of order k. The class of context functions of length k is rich enough to obtain a broad class of Markov chains, including special sparse types given by the notion of a short context. In particular, some context functions $c(.)$ would yield a substantial reduction in the number of parameters compared to a full Markov chain of the same order as the context function. The VLMC's are thus an attractive model class, which is often not much exposed to the curse of dimensionality.

In order to explain our procedure for adaptively selecting and fitting a VLMC, it is most convenient to represent a context function, and hence the set of relevant histories of a VLMC, as a tree. We consider trees with a root node on top, from which the branches are growing downwards, so that every internal node has at most $|\mathbf{X}|$ offsprings. Then, each value of a context function $c(.) : \mathbf{X}^k \to \mathbf{X}^k$ can be represented as a branch (or terminal node) of such a tree. The context $w = c(x_{-k+1}^0)$ is represented by a branch, whose sub-branch on the top is determined by x_0, the next sub-branch by x_{-1} and so on, and the terminal sub-branch by $x_{-\ell(x_0,\ldots,x_{-k+1})+1}$.

Example 2.1 $|\mathbf{X}| = 2$, $k = 3$.
The function

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

can be represented by the tree τ_c,



A 'growing to the left' sub-branch represents the symbol 0 *and vice versa for the symbol* 1.

Note that context trees do not have to be complete, i.e., every internal does not need to have exactly $|\mathbf{X}|$ offsprings (when $|\mathbf{X}| > 2$).

Definition 2.3 *Let $c(.)$ be a context function of a stationary VLMC of order k. The corresponding ($|\mathbf{X}|$-ary) context tree τ and terminal node context tree τ^t are defined as*

$$\tau = \tau_c = \{w; w = c(x_{-k+1}^0),\ x_{-k+1}^0 \in \mathbf{X}^k\},$$
$$\tau^t = \tau_c^t = \{w; w \in \tau_c \text{ and } wu \notin \tau_c \text{ for all } u \in \cup_{m=1}^{\infty}\mathbf{X}^m\}.$$

Definition 2.3 says that only terminal nodes in the tree representation τ are considered as elements of the terminal node context tree τ^t. Clearly, we can reconstruct the context function $c(.)$ from τ_c or τ_c^t. The context tree τ_c is nothing else than the minimal state space of the VLMC P_c. An internal node with $b < |\mathbf{X}|$ offsprings can be implicitly thought to be complete by adding one complementary offspring, lumping the $|\mathbf{X}| - b$ non-present nodes together.

2.1 The context algorithm

Given data $X_1, \ldots, X_n$ from a VLMC P_c, the aim is to find the underlying context function $c(.)$ and an estimate of P_c. We will attack and solve this problem by incorporating ideas from data compression as given by Weinberger et al. (1995). We describe now the algorithm for the aim mentioned above. In the sequel we always make the convention that quantities involving time indices $t \notin \{1, \ldots, n\}$ equal zero (or are irrelevant). Let

$$N(w) = \sum_{t=1}^{n} 1_{[X_t^{t+|w|-1}=w]},\ w \in \mathbf{X}^{\infty}, \tag{2.2}$$

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

$$\hat{p}(w) = N(w)/n,\ \hat{p}(u|w) = \frac{N(uw)}{N(w)},\ w, u \in \mathbf{X}^{\infty}, uw = (\ldots, u_2, u_1, \ldots, w_2, w_1). \tag{2.3}$$

The algorithm below constructs the estimated context tree $\hat{\tau}$ to be the biggest context tree such that

$$\Delta_{wu} = \sum_{x \in \mathbf{X}} \hat{p}(x|wu) \log(\frac{\hat{p}(x|wu)}{\hat{p}(x|w)}) N(wu) \geq K \text{ for all } wu \in \hat{\tau}^t \tag{2.4}$$

with $K = K_n \to \infty$ $(n \to \infty)$ a cut-off to be chosen by the user.

Step 1 Given data $X_1, \ldots, X_n$ taking values in a finite space $\mathbf{X}$, fit a maximal ($|\mathbf{X}|$-ary) context tree, i.e., search for the context function $c_{max}(.)$ with terminal node context tree representation τ_{max}^t, where τ_{max}^t is the biggest tree such that every element (terminal node) in τ_{max}^t has been observed at least twice in the data. This can be formalized as follows:

$$w \in \tau_{max}^t \text{ implies } N(w) \geq 2, \text{ and,}$$
$$\tau_{max}^t \succeq \tau^t, \text{ where } w \in \tau^t \text{ implies } N(w) \geq 2.$$

($\tau_1^t \preceq \tau_2^t$ means: $w \in \tau_1^t \Rightarrow wu \in \tau_2^t$ for some $u \in \cup_{m=0}^{\infty}\mathbf{X}^m$ $(\mathbf{X}^0 = \emptyset)$).
Set $\tau_{(0)}^t = \tau_{max}^t$.

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 to $\tau^t_{(0)}$ and let

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

be 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, i.e., the root node). Replace the context $wu = x^0_{-\ell+1}$ by $w = x^0_{-\ell+2}$ if

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

with $\hat{p}(.)$ and $\hat{p}(.|.)$ as defined in (2.3). Decision about pruning for every terminal node in $\tau^t_{(0)}$ yields a (possibly) smaller tree $\tau_{(1)} \preceq \tau^t_{(0)}$. Let

$$\tau^t_{(1)} = \{w;\ w \in \tau_{(1)} \text{ and } wu \notin \tau_{(1)} \text{ for all } u \in \cup_{m=1}^{\infty} \mathbf{X}^m\}.$$

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

Step 4 If interested in probability sources, estimate the transition probabilities $p(x_1|c(x^0_{-\infty})) = \mathbb{P}[X_1 = x_1|c(X^0_{-\infty}) = c(x^0_{-\infty})]$ by $\hat{p}(x_1|\hat{c}(x^0_{-\infty}))$, where $\hat{p}(.|.)$ is defined as in (2.2).

The pruning in the context algorithm can be viewed as some sort of hierarchical backward selection. Dependence on some values further back in the history should be weaker, so that deep nodes in the context tree are considered, in a hierarchical way, to be less relevant. This hierarchical structure is a clear distinction to the CART algorithm (Breiman et al., 1984), where the tree architecture has no built in time structure.

Consistency for finding an underlying true context function $c_0(.)$ and probability distribution P_{c_0} in a more restrictive setting goes back to Weinberger et al. (1995). We denote by

$$\hat{P}_c \text{ the maximum likelihood (ML) fitted context model on } \tau_c, \tag{2.5}$$

$$\hat{P}_{\hat{c}_0} \text{ the fitted VLMC, induced by Step 4 of the context algorithm.} \tag{2.6}$$

Note that the ML fitted context model on τ_c is given by the estimated transition probabilities $\hat{p}(.|w),\ w \in \tau_c$, where $\hat{p}(.|.)$ is as in (2.3).

For the algorithm described here, consistency even in an asymptotically infinite dimensional setting has been given in Bühlmann and Wyner (1997), where also more detailed descriptions of the context algorithm and cross-connections can be found. An efficiency result is given in Bühlmann (1997). For deriving all these results, we need besides some technical assumptions which we state in section 3 a lower bound for the cut-off value $K_n \sim C \log(n),\ C > (2|\mathbf{X}| + 3)$. In this paper we also develop a strategy for estimating this cut-off K_n as the minimizer of certain risk functions.

3 Risk functions and sub–models

We restrict ourselves now to the following framework: the data X_1^n is a finite realization of a VLMC with context function $c_0(.)$ of finite order k_0 and corresponding context tree τ_{c_0}. We consider sub-models of the true underlying process P_{c_0}, namely the set

$$\{P_c : \ c(.) \text{ a context function with context tree representation } \tau_c \text{ such that } \tau_c \preceq \tau_{c_0}\},$$

where the relation for nestedness of models $\preceq$ is defined in terms of terminal nodes of context trees,

$$\tau_1 \preceq \tau_2 \Longleftrightarrow \left(w \in \tau_1^t \Rightarrow wu \in \tau_2^t \text{ for some } u \in \cup_{m=0}^{\infty} \mathbf{X}^m \ (\mathbf{X}^0 = \emptyset)\right).$$

The problem of (sub-)model selection is studied in terms of two different risk criteria, the final prediction error and the expected Kullback-Leibler information.

3.1 Final prediction error

For a predictor $\hat{Y}_{n+1}$ based on the infinite past $Y_{-\infty}^n$ for a random variable Y_{n+1}, we consider the loss functions

$$L_2(Y_{n+1}, \hat{Y}_{n+1}) = (Y_{n+1} - \hat{Y}_{n+1})^2,$$
$$\delta(Y_{n+1}, \hat{Y}_{n+1}) = 1_{[Y_{n+1} \neq \hat{Y}_{n+1}]}.$$

The L_2 loss can be of interest for ordinal data equipped with some 'Gaussian' scale (quantized Gaussian data) or also for binary data. The δ loss, or zero-one loss, is interesting for categorical data without any order or scale.

The final prediction error (FPE) for the quadratic L_2 loss dates back to Akaike (1969, 1970) and can be generalized in an obvious way for any convex loss function. Let the data X_1^n be a finite realization of the true underlying process P_{c_0} and let $(Y_t)_{t \in \mathbb{Z}}$ be another realization of P_{c_0}, independent of X_1^n. Optimal (theoretical) prediction of Y_{n+1} given the infinite history $Y_{-\infty}^n$ projected on an element of the sub-models $\tau_c \preceq \tau_{c_0}$ with context function $c(.)$ is given by

$$\mathbb{E}_{P_{c_0}}[Y_{n+1}|c(Y_{-\infty}^n)] \text{ for the } L_2 \text{ loss},$$
$$\mathrm{AM}_{P_{c_0}}(c(Y_{-\infty}^n)) = \mathrm{argmax}_{x \in \mathbf{X}} \mathbb{P}_{P_{c_0}}[Y_{n+1} = x|c(Y_{-\infty}^n)] \text{ for the } \delta \text{ loss}.$$

When estimating the theoretical predictors by the data X_1^n, we get

$$\varphi(c(Y_{-\infty}^n), X_1^n) = \begin{cases} \mathbb{E}_{\hat{P}_c}[Y_{n+1}|c(Y_{-\infty}^n)] \text{ for the } L_2 \text{ loss} \\ \mathrm{AM}_{\hat{P}_c}(c(Y_{-\infty}^n)) \text{ for the } \delta \text{ loss} \end{cases}, \tag{3.7}$$

where $\hat{P}_c$ is the estimate in (2.5) based on the data X_1^n.
The predictor $\varphi(.,.)$ could also be defined in terms of the estimated probability measure $\hat{P}_{\hat{c}_0}$ in (2.6). Under appropriate conditions, the two versions are asymptotically equivalent: it is known that for $\tau_c \preceq \tau_{c_0}$, $\mathbb{P}_{\hat{P}_c}[Y_{n+1} = x|c(Y_{-\infty}^n) = w] = \mathbb{P}_{\hat{P}_{\hat{c}_0}}[Y_{n+1} = x|c(Y_{-\infty}^n) = w] + o_P(n^{-1})$ for all $x \in \mathbf{X}$ and all $w \in \tau_c$, cf. Bühlmann and Wyner (1997).
The FPE's for the element P_c with corresponding context tree $\tau_c \preceq \tau_{c_0}$ is then defined as

$$R(\tau_c, P_{c_0}) = \mathbb{E}_{P_{c_0}}[L(Y_{n+1}, \varphi(c(Y_{-\infty}^n), X_1^n))],$$

where $L(.,.) = L_2$ or δ. The general notation $R(.,.)$ indicates that the FPE's are *risk* functions. Specifically,

$$\mathrm{FPE}_{L_2}(\tau_c) = \mathbb{E}_{P_{c_0}}[(Y_{n+1} - \mathbb{E}_{\hat{P}_c}[Y_{n+1}|c(Y_{-\infty}^n)])^2],$$
$$\mathrm{FPE}_{\delta}(\tau_c) = \mathbb{P}_{P_{c_0}}[Y_{n+1} \neq \mathrm{AM}_{\hat{P}_c}(c(Y_{-\infty}^n))].$$

The FPE measures the risk for predicting the observation Y_{n+1} in a new sample Y_1^n when estimation is based on the observed data-set X_1^n. Note that X_1^n is also referred to as training set and Y_1^{n+1} as test set. The following two Theorems describe how the FPE decomposes into an 'oracle part' which is not depending on the model τ_c (when we would know the whole true underlying probability distribution P_{c_0}), a bias part (due to misspecification of the model) and a variance part (due to estimation of the unknown parameters in the model). In the sequel of the paper, we denote by $P(x) = \mathbb{P}_P[X_1^m = x]$ $(x \in \mathbf{X}^m)$ and $P(x|w) = P(xw)/P(w)$ $(x \in \mathbf{X}^{m_1},\ w \in \mathbf{X}^{m_2})$. We then make the following assumptions.

(A1) P_{c_0} satisfies,

$$\sup_{v,w,w'} |p_Z^{(r)}(v,w) - p_Z^{(r)}(v,w')| < 1 - \kappa, \text{ for some } \kappa > 0,$$

where $p_Z^{(r)}(v,w) = \mathbb{P}[Z_r = v|Z_0 = w]$ denotes the r-step transition kernel of the state process $Z_t = c_0(X_0^t x_0^\infty)$, $x_0^\infty = x_0, x_0, \ldots$ $(t \in \mathbb{N}_0)$ with $(X_t)_{t\in\mathbb{Z}} \sim P_{c_0}$.
The definition of Z_t reflects our implicit assumption here that the initial state is padded with elements $x_0 \in \mathbf{X}$, i.e., $Z_0 = w$ means $Z_0 = wx_0^\infty$ so that the next states Z_t $(t > 0)$ are uniquely determined.

(A2) P_{c_0} satisfies

$$\min_{w\in\tau_{c_0}} P_{c_0}(w) > 0,$$
$$\min_{x\in\mathbf{X}, w\in\tau_{c_0}} P_{c_0}(x|w) > 0,$$
$$\min_{wu\in\tau_{c_0}, u\in\mathbf{X}} \sum_{x\in\mathbf{X}} |P_{c_0}(x|wu) - P_{c_0}(x|w)| > 0.$$

Assumption (A1) is a Doeblin-type condition, which has been employed in Bühlmann and Wyner (1997). Assumption (A2) ensures that the VLMC P_{c_0} is not degenerated: the states $w \in \tau_{c_0}$ have all positive probabilities, the transition probabilities are bounded away from zero and the states $wu \in \tau_{c_0}$ are distinguishable from their parent nodes w in the context tree representation.

Theorem 3.1 *Consider a finite realization X_1^n from P_{c_0} satisfying (A1), (A2) and with context tree representation τ_{c_0}. Then, for any element of the sub-models with context function c and corresponding tree representation $\tau_c \preceq \tau_{c_0}$, the following decomposition holds:*

$$FPE_{L_2}(\tau_c) = S + B + V_n,$$
$$S = \mathbb{E}_{P_{c_0}}[(Y_{n+1} - \mathbb{E}_{P_{c_0}}[Y_{n+1}|c_0(Y_{-\infty}^n)])^2],$$
$$B = \left(\mathbb{E}_{P_{c_0}}[Y_{n+1}|c(Y_{-\infty}^n)] - \mathbb{E}_{P_{c_0}}[Y_{n+1}|c_0(Y_{-\infty}^n)]\right)^2,$$
$$V_n = \mathbb{E}_{P_{c_0}}[\left(\varphi(c(Y_{-\infty}^n), X_1^n) - \mathbb{E}_{P_{c_0}}[Y_{n+1}|c(Y_{-\infty}^n)]\right)^2],$$

where $\varphi(c(Y_{-\infty}^n), X_1^n) = \mathbb{E}_{\hat{P}_c}[Y_{n+1}|c(Y_{-\infty}^n)]$ *as in (3.7) and*

$$nV_n - C(\tau_c, P_{c_0}) = o_P(1),$$

$$C(\tau_c, P_{c_0}) = \sum_{w\in\tau_c}\sum_{x_1,x_2\in\mathbf{X}} x_1x_2 \sum_{k=-\infty}^{\infty}\left(P_{c_0}(x_2|w)\mathbb{P}_{P_{c_0}}[X_{-|w|}^0 = x_1w|X_{k-|w|}^k = x_2w] - P_{c_0}(x_1w)\right).$$

The S term is the 'oracle' FPE of order $O(1)$, the B term is the bias term of order $O(1)$ and the V_n term is a penalty term, which behaves asymptotically like $n^{-1}C(\tau_c, P_{c_0})$. The constant $C(\tau_c, P_{c_0})$ is of more complex nature than say the variance term for prediction in an AR(p) model (which behaves as p/n). But by assumption (A2) we still can bound the penalty term linearly in $|\tau_c|$ as

$$C(\tau_c, P_{c_0}) \leq |\tau_c| M(\mathbf{X}, k_0, \kappa),$$

where $M(\mathbf{X}, k_0, \kappa)$ is a constant, depending on the order k_0 of the VLMC P_{c_0} and the value κ in (A1).

For analyzing the FPE$_\delta$ we make the additional rather weak assumption about the uniqueness of the AM$_{P_{c_0}}$,

(B1) For a sub-model P_c with corresponding context tree $\tau_c \preceq \tau_{c_0}$,

$$\min_{w\in\tau_c, k\neq \mathrm{AM}_{P_{c_0}}(w)} |P_{c_0}(\mathrm{AM}_{P_{c_0}}(w)|w) - P_{c_0}(k|w)| > \varepsilon,\ \varepsilon > 0,$$

and denote by $\pi = \min_{w\in\tau_c, x\in\mathbf{X}} P_{c_0}(xw) > 0$.

Note that the fact $\pi > 0$ is implied by assumption (A2).

Theorem 3.2 *Consider a finite realization* X_1^n *from* P_{c_0} *satisfying (A1), (A2) and with context tree representation* τ_{c_0}*. Then, for any element of the sub-models with context function* c *and corresponding tree representation* $\tau_c \preceq \tau_{c_0}$*, satisfying (B1), the following decomposition holds:*

$$\begin{aligned}
&FPE_\delta(\tau_c) = S + B + V_n,\\
&S = \mathbb{P}_{P_{c_0}}[Y_{n+1} \neq AM_{P_{c_0}}(c_0(Y_{-\infty}^n))],\\
&B = \mathbb{P}_{P_{c_0}}[Y_{n+1} \neq AM_{P_{c_0}}(c(Y_{-\infty}^n))] - \mathbb{P}_{P_{c_0}}[Y_{n+1} \neq AM_{P_{c_0}}(c_0(Y_{-\infty}^n))],\\
&V_n = \mathbb{P}_{P_{c_0}}[Y_{n+1} \neq \varphi(c(Y_{-\infty}^n), X_1^n)] - \mathbb{P}_{P_{c_0}}[Y_{n+1} \neq AM_{P_{c_0}}(c(Y_{-\infty}^n))],
\end{aligned}$$

where $\varphi(c(Y_{-\infty}^n), X_1^n) = AM_{\hat{P}_c}(c(Y_{-\infty}^n))$ *as in (3.7) and for* n *sufficiently large,*

$$|V_n| \leq (|\mathbf{X}| + 1)C_1 \exp(-C_2(\kappa)\varepsilon^2\pi^2(n - k_0 + 1)/\log(n - k_0 + 1)),$$

k_0 *the order of* P_{c_0}*,* $C_1 > 0$ *a constant,* $C_2(\kappa) > 0$ *depending only on* κ *in (A1).*

The 'oracle' FPE is again denoted by S being of order $O(1)$, B is the bias term of order $O(1)$. The penalty term V_n decays at least exponentially in n, the size $|\tau_c|$ enters only implicitly into the speed of the exponential decay: larger sub-models have typically smaller values ε and π yielding smaller values $\varepsilon^2\pi^2$ and hence slower, but still exponential decay for the bound of V_n. This suggests that the bias part B is more dominant in FPE$_\delta$ than in FPE$_{L_2}$.

For both types of FPE, Theorems 3.1 and 3.2 show that the S- and B-terms are of constant order $O(1)$, whereas the variance terms V_n decrease as sample size increases.

3.2 Kullback-Leibler information

When considering the goodness of a model in terms of its whole n-dimensional distribution, the Kullback-Leibler information (KLI)

$$\mathrm{KLI}(\tau_c) = I_n(P_{c_0}, \hat{P}_c) = \int_{\mathbf{X}^n} \log\left(\frac{P_{c_0}(y_1^n)}{\hat{P}_c(y_1^n)}\right) dP_{c_0}(y_1^n)$$

measures a loss between the n-dimensional marginals of P_{c_0} and the maximum likelihood estimate $\hat{P}_c$ of a sub-model P_c with context tree representation $\tau_c \preceq \tau_{c_0}$. Similar as with the prediction error, $\hat{P}_c$ is estimated based on the observed data X_1^n, whereas the integration-variable y_1^n can be thought as a new sample (test set). Often one uses as a risk function the expected $\mathrm{KLI}(\tau_c)$,

$$\mathrm{EKLI}(\tau_c) = \mathbb{E}_{P_{c_0}}[I_n(P_{c_0}, \hat{P}_c)]. \tag{3.8}$$

Theorem 3.3 *Consider a finite realization X_1^n from P_{c_0} satisfying (A1), (A2) and with context tree representation τ_{c_0}. Then, for any element of the sub-models with context function c and corresponding tree representation $\tau_c \preceq \tau_{c_0}$, the following decomposition holds:*

$$\begin{aligned}
&KLI(\tau_c)/n = I_n(P_{c_0}, \hat{P}_c)/n = B_n + V_n/n,\\
&B_n = I_n(P_{c_0}, \bar{P}_c)/n,\\
&V_n \Rightarrow \frac{1}{2} Z^T \Sigma(\tau_c, P_{c_0}) Z \ (n \to \infty),
\end{aligned}$$

where $\bar{P}_c$ is the restriction of P_{c_0} on the sub-model structure τ_c, generated by the transition probabilities

$$\bar{P}_c(x|w) = P_{c_0}(xw)/P_{c_0}(w) \text{ for } x \in \mathbf{X},\ w \in \tau_c,$$

$Z \sim \mathcal{N}_{D(\tau_c)}(0, I)$, $D(\tau_c) = |\tau_c|(|\mathbf{X}| - 1)$ the dimension of the sub-model, and $\Sigma(\tau_c, P_{c_0})$ a non-degenerate $D(\tau_c) \times D(\tau_c)$ matrix, depending on the sub-model structure τ_c and the underlying process P_{c_0}.

The B_n term is a bias part of the constant order $O(1)$ due to misspecification of the model, and V_n/n is a penalty term of the order $O_P(n^{-1})$. More insight about the matrix $\Sigma(\tau_c, P_{c_0})$ can can be obtained from the proof in section 7.

Remark 3.1. Tong (1975) derives the limiting χ^2-distribution of 2 times the V_n term for a full Markov chain. Although not explicitly pointed out, this only holds for $\tau_c = \tau_{c_0}$ being the true model: then $\Sigma(\tau_{c_0}, P_{c_0}) = I_{D(\tau_{c_0})}$ and the limiting distribution of V_n equals $\chi^2_{D(\tau_{c_0})}/2$. The limiting distribution of V_n in general is connected to the derivation of the TIC criterion (Takeuchi, 1976), see also Shibata (1989, section 2): this approach accounts for the effect that the true model is generally not equal to the fitted model.

4 A bootstrap method for estimating risk functions

An often used approach to estimate the various risk functions in section 3 is given by estimating the different terms in Theorems 3.1-3.3. Criteria like AIC, BIC, TIC, cf. Shibata (1989), are aiming to minimize a criterion function 'goodness of fit + penalty'. They essentially estimate the unknown asymptotic values in Theorems 3.1-3.3: the $(S+B)$-terms by a goodness of fit statistic, i.e., residual sum of squares in the Gaussian case, and the V_n-terms by different strategies. More recently, the idea of bootstrap in model selection has been pursued, but mainly for bias correction in the estimation of the penalty term, cf. Efron (1983, 1986), Cavanaugh and Shumway (1997), Shibata (1997) clarifies about different bootstrap strategies for bias corrections.

We propose here a model selection approach for the dependent setting with VLMC's which is entirely driven by a bootstrap scheme, rather than only making a bias correction via resampling for estimation of a penalty term. This seems more appealing than combining estimation of $(S+B)$-terms, V_n-terms and bias correction for the V_n-terms. Also, resampling schemes are potentially able to pick up not only a bias but also higher order cumulants. In principle, estimation of (conditional) prediction errors (but not risks in the sense of an expected prediction error) could also be done with some cross-validation technique for dependent data. However, cross-validation estimates are usually highly variable, cf. Efron (1983), and thus not very accurate.

Below is the general principle for estimating a risk function of P_c with structure $\tau_c \preceq \tau_{c_0}$, being a sub-model of the true underlying process P_{c_0}. Assume that we have given data $X_1, \ldots, X_n$.

Step 1 Fit with the context algorithm in section 2.1 a VLMC $\hat{P}_{\hat{c}_0}$ as in (2.6).

Step 2 For a context model with structure $\tau_c \preceq \tau_{c_0}$, compute the bootstrap risk functions,

$$\begin{aligned}
&\mathrm{FPE}^*(\tau_c) = \mathbb{E}_{\hat{P}_{\hat{c}_0}}[L(Y^*_{n+1}, \varphi(c((Y^*)_1^n), (X^*)_1^n)|X_1^n],\ L = L_2,\ \delta,\\
&\mathrm{KLI}^*(\tau_c) = I_n(\hat{P}_{\hat{c}_0}, \hat{P}^*_c),
\end{aligned}$$

where $\varphi(.,.)$ is as in (3.7) and

$$\begin{aligned}
(Y^*)_1^{n+1} &\sim \hat{P}_{\hat{c}_0} \circ \pi^{-1}_{1,\ldots,n+1},\\
(X^*)_1^{n+1} &\sim \hat{P}_{\hat{c}_0} \circ \pi^{-1}_{1,\ldots,n},
\end{aligned} \tag{4.9}$$

with $(Y^*)_1^{n+1}$ and $(X^*)_1^n$ being independent finite realizations of the fitted model $\hat{P}_{\hat{c}_0}$ in (2.6) based on the data X_1^n, and $\pi_{1,\ldots,m}$ $(m \in \mathbb{N})$ the coordinate function. The estimate

$$\hat{P}^*_c = T_c((X^*)_1^n) \tag{4.10}$$

is the plug-in version of the ML fitted context model $\hat{P}_c = T_c(X_1^n)$ on τ_c, as in (2.5).

The bootstrap $\mathrm{FPE}^*(\tau_c)$ is then directly used as an estimate of the true $\mathrm{FPE}(\tau_c)$, the bootstrap $\mathrm{KLI}^*(\tau_c)$ is a random variable depending on $(X^*)_1^n$ (given the original sample X_1^n): often, one is interested in $\mathrm{EKLI}^*(\tau_c) = \mathbb{E}_{\hat{P}_{\hat{c}_0}}[I_n(\hat{P}_{\hat{c}_0}, \hat{P}^*_c)|X_1^n]$ as an estimate of $\mathrm{EKLI}(\tau_c)$

as defined in (3.8). In practice, the expectations with respect to $\hat{P}_{\hat{c}_0}$ are evaluated via Monte-Carlo. Minimization of such estimated risks over all (or some) sub-models $\tau_c \preceq \tau_{c_0}$ of the true underlying VLMC P_{c_0} yields in theory the estimated optimal (or sub-optimal) model. The initial estimate $\hat{P}_{\hat{c}_0}$ serves as an approximation for the true underlying process P_{c_0}.

Theorem 4.1 *Assume the situation and notation in Theorem 3.1. Moreover, suppose that the cut-off $K_n > (2|\mathbf{X}|+3)\log(n)$ in Step 2 of the context algorithm for constructing the estimate $\hat{P}_{\hat{c}_0}$. Then,*

$$\begin{aligned}
&FPE^*_{L_2}(\tau_c) = S^* + B^* + V^*_n,\\
&S^* = S + o_P(1)\ (n \to \infty),\\
&B^* = B + o_P(1)\ (n \to \infty),\\
&V^*_n = V_n + o_P(n^{-1})\ (n \to \infty).
\end{aligned}$$

The quantities S^*, B^* and V^*_n are the plug-in versions of S, B and V_n, respectively with $\hat{P}_{\hat{c}_0}$ instead of P_{c_0} and $\hat{c}_0$ instead of c_0.

Theorem 4.2 *Assume the situation and notation in Theorem 3.2. Moreover, suppose that the cut-off $K_n > (2|\mathbf{X}|+3)\log(n)$ in Step 2 of the context algorithm for constructing the estimate $\hat{P}_{\hat{c}_0}$. Then,*

$$\begin{aligned}
&FPE^*_{\delta}(\tau_c) = S^* + B^* + V^*_n,\\
&S^* = S + o_P(1)\ (n \to \infty),\\
&B^* = B + o_P(1)\ (n \to \infty),\\
&V^*_n = O_P(\exp(-Cn))\ (n \to \infty),\ C > 0 \text{ a constant.}
\end{aligned}$$

The quantities S^*, B^* and V^*_n are the plug-in versions of S, B and V_n, respectively, with $\hat{P}_{\hat{c}_0}$ instead of P_{c_0} and $\hat{c}_0$ instead of c_0.

Theorem 4.3 *Assume the situation and notation in Theorem 3.3. Moreover, suppose that the cut-off $K_n > (2|\mathbf{X}|+3)\log(n)$ in Step 2 of the context algorithm for constructing the estimate $\hat{P}_{\hat{c}_0}$. Then,*

$$\begin{aligned}
&KLI^*(\tau_c)/n = I_n(\hat{P}_{\hat{c}_0}, \hat{P}^*_c)/n = B^*_n + V^*_n/n,\\
&B^*_n = B_n + o_P(1)\ (n \to \infty),\\
&V^*_n \Rightarrow \text{ (limiting distribution of } V_n\text{) in probability as } (n \to \infty).
\end{aligned}$$

The quantities B^*_n and V^*_n are the plug-in versions of B_n and V_n, respectively, with $\hat{P}_{\hat{c}_0}$ instead of P_{c_0} and $\hat{c}_0$ instead of c_0.

Remark 4.1. Theorems 4.1-4.3 describe the consistency of the bootstrap risk estimator, even for the higher order V_n-terms. Consistency for the V_n terms is important for high-dimensional parameter spaces, here given by the number $D(\tau_c)$: if $D(\tau_c)$ is large, then the V_n-terms are typically not that much negligible compared to the $(S+B)$-terms.

Remark 4.2. Using $\mathrm{EKLI}^*(\tau_c) = \mathbb{E}_{\hat{P}_{\hat{c}_0}}[\mathrm{KLI}^*(\tau_c)|X_1^n]$ as a criterion for model selection is not equivalent to AIC. In general, the term penalizing large models in $\mathrm{EKLI}^*(\tau_c)$ is

not converging to (the wrong constant) $D(\tau_c)/2$ which would correspond to the equivalent penalty term $2D(\tau_c)$ in AIC. In our set-up, AIC is generally not a consistent criterion for minimizing EKLI(τ_c).

Remark 4.3. It has been pointed out by Efron (1983) that estimation of a prediction error with the nonparametric bootstrap in the i.i.d. case has a potential to underestimate. But the informal distance arguments, leading also to Efron's .632 estimator, lack any heuristics here because our resampling is based on a (semi-)*parametrically* estimated VLMC $\hat{P}_{\hat{c}_0}$.

4.1 Tuning the context algorithm

We denote in the sequel by

$$R(\tau_c) = \begin{cases} \mathrm{FPE}_{L_2}(\tau_c) \\ \mathrm{FPE}_{\delta}(\tau_c) \\ \mathrm{EKLI}(\tau_c) \end{cases}$$

one of the different risk functions in section 3 (thereby notationally neglecting the dependence on P_{c_0}). Even when we would know the risk function $R(\tau_c)$ for all sub-models $\tau_c \preceq \tau_{c_0}$, the search over all these sub-models can be computationally infeasible. We focus here on the problem of finding the best sub-model among the models produced by the context algorithm.

Denote by $\hat{\tau}_0 = \tau^t_{max}$ the maximal context tree as in Step 1 of the context algorithm in section 2.1. By successively increasing the cut-off value K in Step 2 of the context algorithm, we get a finite sequence of nested context tree estimates,

$$\hat{\tau}_0 \succ \hat{\tau}_1 \succ \ldots \succ \hat{\tau}_{\hat{m}-1} \succ \tau_{\hat{m}} = \tau_{root}, \tag{4.11}$$

where τ_{root} is the root corresponding to independence.
Note that the trees $\hat{\tau}_k$ $(0 \leq k \leq \hat{m}-1)$ and $\hat{m}$ depend on the data X_1^n. We can thus think of a cut-off K as a selection rule,

$$K : X_1^n \to \hat{\tau}_K, \ \hat{\tau}_K \in \{\hat{\tau}_0, \ldots, \hat{\tau}_{\hat{m}-1}, \tau_{root}\}. \tag{4.12}$$

What we want is to minimize an overall risk $R'(K)$ over cut-off parameters (or selection rules) K, with $R'(.)$ now also taking into account the randomness of the tree $\hat{\tau}_K$. Note that the randomness comes in by the context algorithm and would also be present, even if risk functions for fixed models τ_c would be completely known. Denote by $\hat{c}_K$ the estimated context function with corresponding tree representation $K(X_1^n) = \hat{\tau}_K$ as in (4.12). We define the overall risk $R'(.)$ as

$$R'(K) = \begin{cases} \mathbb{E}_{P_{c_0}}[L(Y_{n+1}, \varphi_K(\hat{c}_K(Y_{-\infty}^n), X_1^n))] \text{ for FPE with } L = L_2, \ \delta \\ \mathbb{E}_{P_{c_0}}[I_n(P_{c_0}, \hat{P}_{\hat{c}_K})] = \mathbb{E}_{P_{c_0}}[\int_{\mathbf{X}^n} \log\left(\frac{P_{c_0}(y_1^n)}{\hat{P}_{\hat{c}_K}(y_1^n)}\right) dP_{c_0}(y_1^n)] \text{ for EKLI} \end{cases}, \tag{4.13}$$

where

$$\varphi_K(\hat{c}_K(Y_{-\infty}^n), X_1^n) = \begin{cases} \mathbb{E}_{\hat{P}_{\hat{c}_K}}[Y_{n+1}|\hat{c}_K(Y_{-\infty}^n)])^2 \text{ for the } L_2 \text{ loss} \\ \mathrm{AM}_{\hat{P}_{\hat{c}_K}}(\hat{c}_K(Y_{-\infty}^n)) \text{ for the } \delta \text{ loss} \end{cases},$$

and $\hat{P}_{\hat{c}_K}$ as in (2.6), but now with a notation emphasizing the dependence on the cut-off K.
The optimal cut-off is then

$$K_{opt} = \text{argmin}_K R'(K). \tag{4.14}$$

Estimation of $R'(.)$ is again proposed by a bootstrap scheme. Let $\hat{c}_K^*$ be the bootstrap version with corresponding context tree $\hat{\tau}_K^* = K((X^*)_1^n)$, $K(.)$ as in (4.12). The bootstrap estimation of the overall risk $R'(K)$ is then pursued similarly as in the previous section by the plug-in principle.

Step 1 For a cut-off K_0, fit a VLMC $\hat{P}_{\hat{c}_{K_0}}$ as in (2.6).

Step 2 Compute the bootstrap risk functions

$$\text{FPE}^*(K) = \mathbb{E}_{\hat{P}_{\hat{c}_{K_0}}}[L(Y_{n+1}^*, \varphi_K(\hat{c}_K^*((Y^*)_1^n), (X^*)_1^n)|X_1^n],\ L = L_2,\ \delta,$$
$$\text{EKLI}^*(K) = \mathbb{E}_{\hat{P}_{\hat{c}_{K_0}}}[I_n(\hat{P}_{\hat{c}_{K_0}}, \hat{P}_{\hat{c}_K^*}^*)|X_1^n],$$

where $(Y^*)_1^{n+1}$, $(X^*)_1^n$ are as in (4.9) but with $\hat{P}_{\hat{c}_K}$ replacing the notation $\hat{P}_{\hat{c}_0}$.

The data-driven cut-off values are then defined as

$$\hat{K} = \text{argmin}_K \text{FPE}^*(K) \text{ or } \hat{K} = \text{argmin}_K \text{EKLI}^*(K). \tag{4.15}$$

Rigorous mathematical results for $\text{FPE}^*(K)$, $\text{EKLI}^*(K)$ or $\hat{K}$ in (4.15) are difficult to obtain due to the randomness of a context function $\hat{c}_K$ for a given cut-off K. When treating $\hat{c}_K$ as fixed and hence incorrectly ignoring its stochastic nature, we are back in the set-up of Theorems 4.1 - 4.3. It is an open question how to fill this gap in theory. The performance of the algorithmic implementation for finite sample sizes is investigated in section 5.

4.1.1 Relation to optimal pruned subtrees

Assume that we know the risk function $R(\tau_c)$ for all fixed sub-model structures $\tau_c \preceq \tau_{c_0}$. Optimality within the sequence of nested trees $\{\hat{\tau}_k\}_k$ in (4.11) then motivates the definition

$$\tilde{\tau}_{opt} = \tilde{\tau}_{opt}(X_1^n) = \text{argmin}_{\hat{\tau}_k} R(\hat{\tau}_k).$$

The tree $\tilde{\tau}_{opt}$, which *depends on the data*, is called the 'optimal pruned sub-tree' with respect to the risk function $R(.)$, cf. Breiman et al. (1984, chapters 3.3-3.4, 10). When the risk $R(.)$ is unknown, we can replace it by some estimate, in our case by e.g. the bootstrap estimate $R^*(.)$.

However, the tree $\tilde{\tau}_{opt}$ might not be optimal with respect to some overall risk $R'(.)$ as in (4.13), treating $\hat{\tau}_k$ as random. When looking at $R'(.)$, we again have to consider a selection rule, say

$$T : X_1^n \to \hat{\tau}_T \in \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{m}-1}, \tau_{root}\},$$

with $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{m}-1}, \tau_{root}\}$ as in (4.11). But such a rule T must coincide with the selection rule given by the cut-off K in (4.12). Thus, our algorithmic implementation in section 4.1 can be interpreted as optimal subtree pruning with respect to some overall risk $R'(.)$.

5 Numerical examples

We study our method for tuning the context algorithm on some simulations for two different models.

5.1 Computational implementation

Approximate calculation of $\mathrm{FPE}^*(K)$ in Step 2 of the algorithm in section 4.1 can be done via Monte Carlo with B replicates in a quite standard way. We always use here $B = 100$.

1. Generate for $i = 1, \ldots, B$,

$$\mathbf{X}_i^* = (X_{i,1}^*, \ldots, X_{i,n}^*) \sim \hat{P}_{\hat{c}_{K_0}} \circ \pi_{1,\ldots,n}^{-1},$$
$$\mathbf{Y}_i^* = (Y_{i,1}^*, \ldots, Y_{i,n}^*, Y_{i,n+1}^*) \sim \hat{P}_{\hat{c}_{K_0}} \circ \pi_{1,\ldots,n+1}^{-1},$$

   where $\mathbf{X}_i^*$, $\mathbf{Y}_j^*$ independent for all i, j, $\mathbf{X}_i^*$, $\mathbf{X}_j^*$ independent for $i \neq j$, $\mathbf{Y}_i^*$, $\mathbf{Y}_j^*$ independent for $i \neq j$.

2. For $i = 1, \ldots, B$, compute $\hat{c}_{i,K}^*$, based on $\mathbf{X}_i^*$ and given by the context tree representation $\tau_{\hat{c}_{i,K}^*} = K(\mathbf{X}_i^*)$, with $K(.)$ being the selection-rule (cut-off) as given in (4.12). Then calculate $\varphi_K(\hat{c}_{i,K}^*((\mathbf{Y}_i^*)_1^n), \mathbf{X}_i^*)$ and set

$$L_i = L(Y_{i,n+1}^*, \varphi_K(\hat{c}_{i,K}^*((\mathbf{Y}_i^*)_1^n), \mathbf{X}_i^*)).$$

3. Use $B^{-1} \sum_{i=1}^B L_i$ as an approximation for $\mathrm{FPE}^*(K)$.

Instead of $\mathrm{EKLI}(K)$ as a risk for selection of K, we consider the negative expected log-likelihood function (NELL), which is equivalent for the purpose of minimization, but computationally cheaper,

$$\mathrm{NELL}(K) = -\int_{\mathbf{X}^n} \log(\hat{P}_{\hat{c}_K}(y_1^n)) dP_{c_0}(y_1^n),$$
$$\mathrm{ENELL}(K) = \mathbb{E}_{P_{c_0}}[\mathrm{NELL}(K)]. \tag{5.16}$$

The approximate calculation of $\mathrm{ENELL}^*(K)$, analogous as for $\mathrm{EKLI}^*(K)$ in Step 2 of the algorithm in section 4.1, can be done without integrating over $\mathbf{X}^n$. We proceed again by Monte Carlo with B replicates,

1. For $i = 1, \ldots, B$, generate analogously as in Step 2 of the algorithm in section 4.1,

$$\mathbf{X}_i^* = (X_{i,1}^*, \ldots, X_{i,n}^*),$$
$$\mathbf{Y}_i^* = (Y_{i,1}^*, \ldots, Y_{i,n}^*).$$

2. For $i = 1, \ldots, B$, compute $\hat{c}_{i,K}^*$, based on $\mathbf{X}_i^*$ and given by the context tree representation $\tau_{\hat{c}_{i,K}^*} = K(\mathbf{X}_i^*)$, and then calculate

$$E_i = -\log(\hat{P}_{\hat{c}_{i,K}^*}^*(\mathbf{Y}_i^*)),$$

   where $\hat{P}_{\hat{c}_{i,K}^*}^*$ is given in (4.10), based on $\mathbf{X}_i^*$.

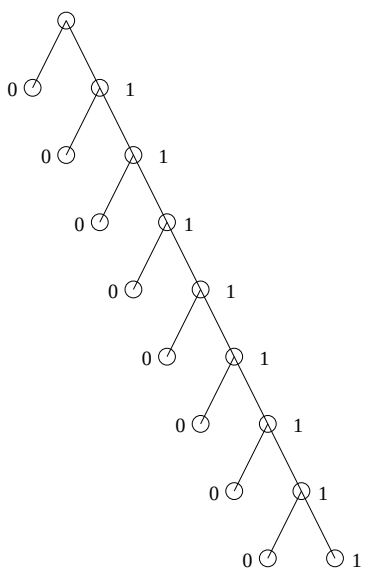
3. Use $B^{-1}\sum_{i=1}^{B} E_i$ as an approximation for $\mathrm{ENELL}^*(K)$.

We use again $B = 100$. It is interesting to note that it is sufficient to compute for every replicate set with label i only one value E_i instead of an n-dimensional integral. The one *single* Monte Carlo iteration over the index set $i = 1, \ldots, B$ takes care about the integration in $\mathrm{NELL}^*(K)$, compare with formula (5.16), as well as of the expectation $\mathbb{E}_{\hat{P}_{\hat{c}_{K_0}}}[\mathrm{NELL}^*(K)]$.

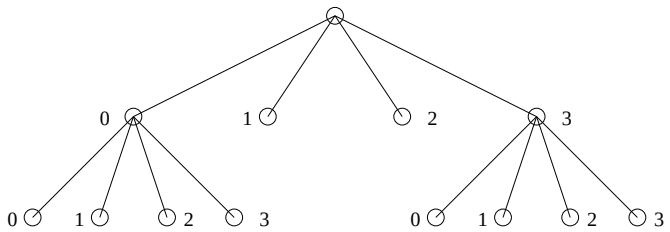
5.2 Simulations

We consider the VLMC's P_{c_0}, represented by the following context trees. The tuple of values at a terminal node w represents the transition probabilities $(P_{c_0}(0|w), \ldots, P_{c_0}(|\mathbf{X}|-1, w))$.

(M1) Binary VLMC of order 8 ($\mathbf{X} = \{0, 1\}$).



(M2) 4-ary VLMC of order 2 ($\mathbf{X} = \{0, 1, 2, 3\}$).



We consider estimation of the different overall risks $R'(K)$ (FPE_{L_2}, FPE_δ and ENELL as in (5.16)) for different initial cut-off values K_0 and the risks $R'(\hat{K})$ when using the estimated cut-off parameter $\hat{K}$ in (4.15). The sample sizes in this study are $n = 200$ and $n = 1000$.

The estimated risks $\hat{R}'(K)$ are computed as described in section 5.1 based on 100 bootstrap replicates. We choose as initial cut-offs K_0 the values $\chi^2_{|\mathbf{X}|-1;0.9}/2$ and $\chi^2_{|\mathbf{X}|-1;0.8}/2$, respectively: the $\chi^2/2$ quantiles, as the limiting quantiles for one log-likelihood ratio test when considering to prune one terminal node in the context algorithm, serve as a good platform for the magnitude of a cut-off.

Figures 5.2 and 5.2 show a sample version of $\mathbb{E}_{P_{c_0}}[\hat{R}'(K)]$, based on 100 simulations of the true process P_{c_0}. The cut-off values $\hat{K}$ in (4.15) are estimated for every individual realization, based on 100 bootstrap replicates. A sample version of $\mathbb{E}_{P_{c_0}}[R'(\hat{K})]$ is then computed over 100 simulations. We compare this with sample versions of $R'(K_{opt}) = min_K R'(K)$ and with sample versions of R_{oracle}, i.e., the risk when knowing the true process P_{c_0} *. All the sample versions are based on 100 simulations of the true process P_{c_0}.

Results are given in Tables 5.2 - 5.2 and graphically displayed in Figures 5.2 - 5.2. The risk function ENELL is always standardized by the factor n^{-1}.
We can summarize as follows.

*The oracle FPE is the risk for the theoretically optimal predictor $\mathbb{E}_{P_{c_0}}[Y_{n+1}|c_0(Y_{-\infty}^n)]$ or $\text{AM}_{P_{c_0}}(c_0(Y_{-\infty}^n))$, respectively. The oracle ENELL is $-\mathbb{E}_{P_{c_0}}[\log(P_{c_0}(Y_1^n))]$.

model, risk, K_0	$\mathbb{E}[R'(\hat{K})]$	$\mathbb{E}[R'(\hat{K})]/R'(K_{opt})$	$\mathbb{E}[R'(\hat{K})]/R_{oracle}$
(M1), FPE_{L_2}, $K_0 = 1.35$	0.21 (0.02)	1.17	1.21
(M1), FPE_{L_2}, $K_0 = 0.82$	0.23 (0.03)	1.28	1.33
(M2), FPE_{δ}, $K_0 = 3.13$	0.21 (0.04)	1.00	1.11
(M2), FPE_{δ}, $K_0 = 2.32$	0.22 (0.04)	1.05	1.16
(M2), ENELL/n, $K_0 = 3.13$	0.87 (0.01)	1.02	1.20
(M2), ENELL/n, $K_0 = 2.32$	0.87 (0.01)	1.02	1.20

Table 5.1: Risks for sample size $n = 200$.

model, risk, K_0	$\mathbb{E}[R'(\hat{K})]$	$\mathbb{E}[R'(\hat{K})]/R'(K_{opt})$	$\mathbb{E}[R'(\hat{K})]/R_{oracle}$
(M1), FPE_{L_2}, $K_0 = 1.35$	0.20 (0.02)	1.11	1.14
(M1), FPE_{L_2}, $K_0 = 0.82$	0.22 (0.03)	1.26	1.23
(M2), FPE_{δ}, $K_0 = 3.13$	0.20 (0.04)	1.11	1.11
(M2), FPE_{δ}, $K_0 = 2.32$	0.21 (0.04)	1.17	1.17
(M2), ENELL/n, $K_0 = 3.13$	0.742 (0.001)	1.01	1.03
(M2), ENELL/n, $K_0 = 2.32$	0.754 (0.003)	1.02	1.05

Table 5.2: Risks for sample size $n = 1000$.

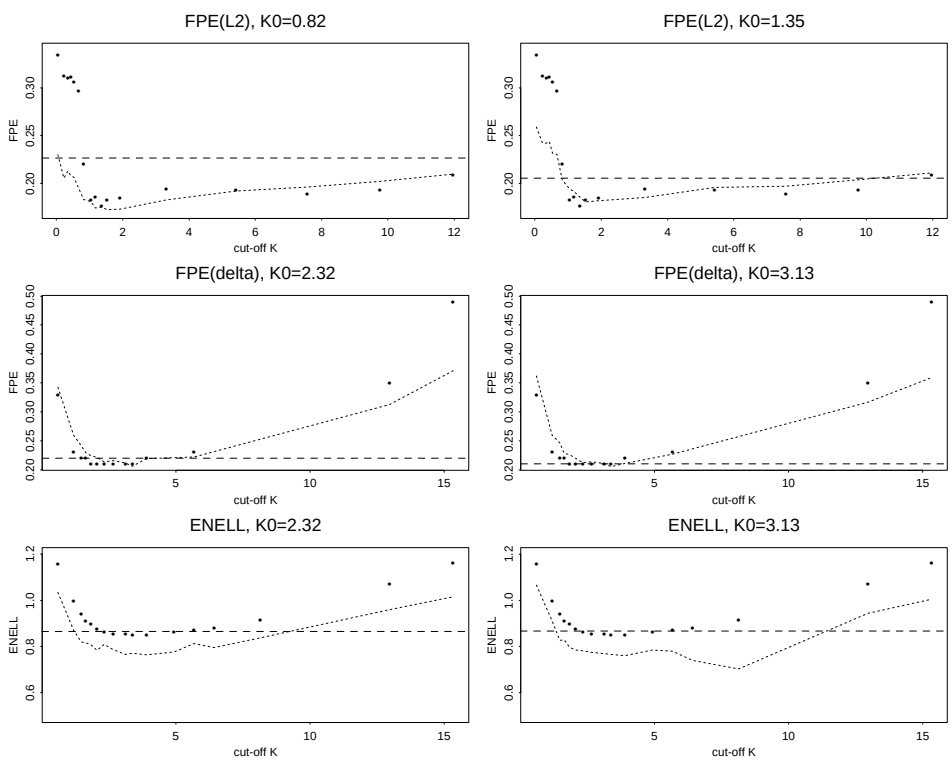


Figure 5.1: Risks for sample size $n = 200$. Model (M1) for FPE_{L_2}, model (M2) for FPE_{δ} and ENELL, respectively. Dots: $R'(K)$; dotted line: $\mathbb{E}[\hat{R}'(K)]$; dashed line: $\mathbb{E}[R'(\hat{K})]$.

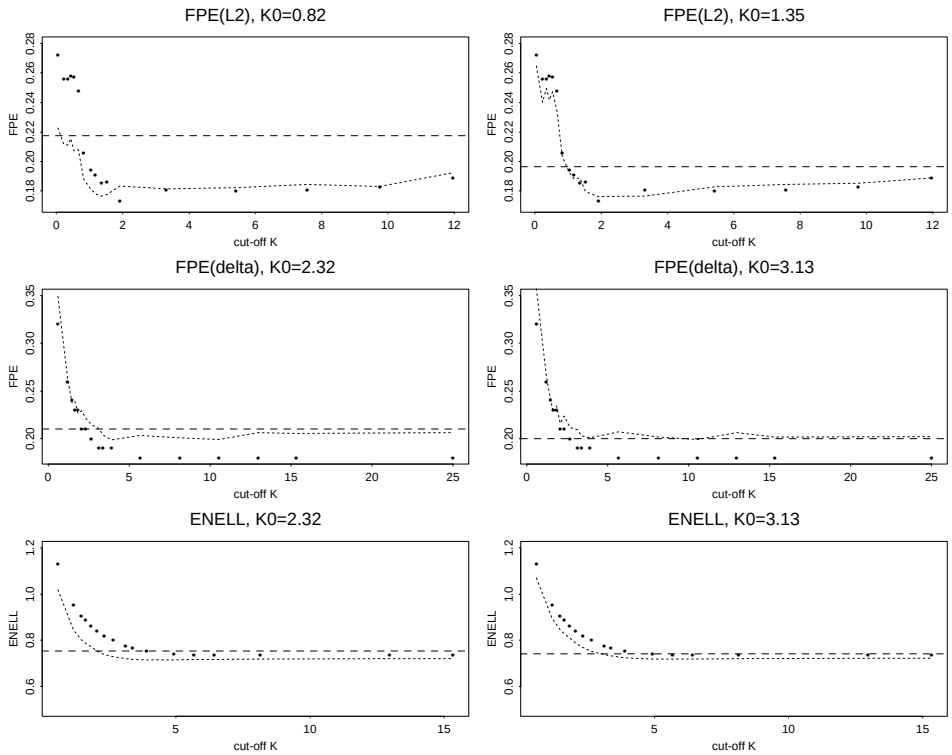


Figure 5.2: Risks for sample size $n = 1000$. Model (M1) for FPE_{L_2}, model (M2) for FPE_δ and ENELL, respectively. Dots: $R'(K)$; dotted line: $\mathbb{E}[\hat{R}'(K)]$; dashed line: $\mathbb{E}[R'(\hat{K})]$.

1. The increase in risk by using $\hat{K}$ instead of the theoretically optimal K_{opt} is biggest in the cases [(M1), FPE_{L_2}], at most 28% for $n = 200$ and 26% for $n = 1000$. In the best cases, the loss is 0% for $n = 200$ and 1% for $n = 1000$.

2. The ratio $\mathbb{E}[R'(\hat{K})]/R'(K_{opt})$ does not necessarily improve with larger sample size. This is due to the fact that the gain for $R'(K_{opt})$ with larger sample size can dominate the gain of $\mathbb{E}[R'(\hat{K})]$ with increasing sample size. But $\mathbb{E}[R'(\hat{K})]$ always improves with increasing sample size, up to the non-significant difference in case [(M2), FPE_δ, $K_0 = 2.32$] due to the finite averaging over 100 simulations.

3. The sensitivity on the initial cut-off K_0 is not very big. The most sensitive cases are [(M1), FPE_{L_2}], which are also the most difficult cases in terms of performance.

4. Figures 5.2 and 5.2 show that even if estimation of $R'(.)$ has a substantial bias, i.e. $|\mathbb{E}[\hat{R}'(K)] - R'(K)|$ large, the substituted minimizers of $R'(.)$ and $\mathbb{E}[\hat{R}'(.)]$ yield rather similar risks, i.e., $|R'(\text{argmin}_K R'(K)) - R'(\text{argmin}_K \mathbb{E}[\hat{R}'(K)])|$ small. This explains visually that using $\hat{K}$ instead of K_{opt} works reasonably well.

6 Conclusions

We have shown in section 3 the asymptotic behavior of different risk functions for submodels in the class of finite space variable length Markov chains. The choice of the loss

function matters and asymptotic equivalence among different risks is not true in general. Depending on the application and pre-knowledge, the flexibility of choosing loss functions can be important.

A semiparametric type bootstrap scheme is then proposed in section 4. It is shown to be asymptotically valid for estimating risks, even for higher order variance parts, and it can then be used for model selection among variable length Markov chains. The bootstrap approach is attractive since it is generally applicable for various loss functions, and model selection can then be done with an optimality focus for specific aims, such as predicting a new observation or estimating the underlying n-dimensional distribution. In the special case of estimating the order of full Markov chains, our methodology also improves the AIC criterion which has been proposed in the past.

From the abstract semiparametric bootstrap principle for estimating risks in section 4 we also gain insight how to choose the cut-off parameter K in the context algorithm, see section 4.1. The problem of tuning the context algorithm is very important for practical applications. The idea is somewhat related to optimal tree pruning in Breiman et al. (1984, chapter 11.7) for CART with independent observations, but our approach takes the randomness of a pruned tree into account. As in model selection mentioned above, our method allows again a tuning which is tailored towards some specific aims, which can be chosen by the user via an appropriate loss function. A simulation study in section 5 confirms the usefulness and robustness of our tuning proposal.

7 Proofs

We usually suppress the index P_{c_0} for moments or probabilities with respect to the measure P_{c_0}.

Proof of Theorem 3.1: The decomposition $\mathrm{FPE}_{L_2}(\tau_c) = S + B + V_n$ follows by the fact that

$$\begin{aligned}
&\mathbb{E}_{P_{c_0}}[Y_{n+1} - \mathbb{E}[Y_{n+1}|c(Y_{-\infty}^n)]|X_1^n, c(Y_{-\infty}^n)] = 0 \ a.s.(P_{c_0}),\\
&\mathbb{E}_{P_{c_0}}[Y_{n+1} - \mathbb{E}[Y_{n+1}|c_0(Y_{-\infty}^n)]|X_1^n, c_0(Y_{-\infty}^n)] = 0 \ a.s.(P_{c_0}).
\end{aligned}$$

It remains to analyze the V_n part. Denote by

$$\begin{aligned}
\hat{\xi} &= \hat{\xi}(c(Y_{-\infty}^n)) = \varphi(c(Y_{-\infty}^n), X_1^n) = \mathbb{E}_{\hat{P}_c}[Y_{n+1}|c(Y_{-\infty}^n)],\\
\xi &= \xi(c(Y_{-\infty}^n)) = \mathbb{E}_{P_{c_0}}[Y_{n+1}|c(Y_{-\infty}^n)].
\end{aligned}$$

Then,

$$\begin{aligned}
V_n &= \mathbb{E}[\mathbb{E}[(\hat{\xi} - \xi)^2|c(Y_{-\infty}^n)]]\\
&= \mathbb{E}[Var(\hat{\xi}|c(Y_{-\infty}^n))] + \mathbb{E}[(\mathbb{E}[\hat{\xi}|c(Y_{-\infty}^n)] - \xi)^2] = I_n + II_n. \qquad (7.17)
\end{aligned}$$

We first show that II_n is asymptotically negligible. Fix $w = c(Y_{-\infty}^n)$ and note that by assumption (A2) $P_{c_0}(w) > 0$. Then, with $n' = n - |w|$ and for $x \in \mathbf{X}$,

$$\begin{aligned}
&\hat{P}_c(x|w) = \frac{N(xw)}{N(w)} = \frac{n'^{-1}N(xw)}{P_{c_0}(w)}\\
- &\frac{n'^{-1}N(xw)}{P_{c_0}^2(w)}(n'^{-1}N(w) - P_{c_0}(w)) + 2\frac{n'^{-1}N(xw)}{\tilde{P}^3(w)}(n'^{-1}N(w) - P_{c_0}(w))^2, \qquad (7.18)
\end{aligned}$$

where $|\tilde{P}(w) - P_{c_0}(w)| \leq |n'^{-1}N(w) - P_{c_0}(w)|$, and $N(.)$ as in (2.2).
By assumption (A1), which ensures the geometric ϕ-mixing property, cf. Doukhan (1994) or see also our remark 7.1, we get

$$n^{1/2}(n'^{-1}N(w) - P_{c_0}(w)) \Rightarrow \mathcal{N}(0, \sigma^2(w)),$$
$$\sigma^2(w) = \sum_{k=-\infty}^{\infty} Cov(1_{[X_0^{m-1}=w]}, 1_{[X_k^{k+m-1}=w]}), \ m = |w|, \tag{7.19}$$

and

$$nCov(n'^{-1}N(xw), n'^{-1}N(w)) \to \tau^2(xw),$$
$$\tau^2(xw) = \sum_{k=-\infty}^{\infty} Cov(1_{[X_0^{m}=xw]}, 1_{[X_k^{k+m-1}=w]}), \ m = |w|. \tag{7.20}$$

Using (7.19), (7.20) and uniform integrability of $\frac{n'^{-1}N(xw)}{\tilde{P}^3(w)} n(n'^{-1}N(w) - P_{c_0}(w))^2$ (this can be shown by using $\tilde{P}(w) > 0$ $a.s.(P_{c_0})$, $0 \leq n'^{-1}N(xw) \leq 1$ and by the geometric ϕ-mixing property of P_{c_0}, implied by (A1), together with the boundedness of indicator functions) we get

$$n\mathbb{E}[\hat{P}_c(x|w) - P_{c_0}(x|w)|w] = -\frac{1}{P_{c_0}^2(w)}\tau^2(xw) + 2\frac{P_{c_0}(x|w)}{P_{c_0}^2(w)}\sigma^2(w) + o(1). \tag{7.21}$$

With (7.21) and the finiteness of τ_c we get

$$II_n = \mathbb{E}[(\mathbb{E}[\hat{\xi}|c(Y_{-\infty}^n)] - \xi)^2] = O(n^{-2}). \tag{7.22}$$

For the variance part I_n we write for fixed $w = c(Y_{-\infty}^n)$,

$$nVar(\hat{\xi}|w) = \sum_{x_1,x_2 \in \mathbf{X}} x_1 x_2 nCov(\frac{N(x_1 w)}{N(w)}, \frac{N(x_2 w)}{N(w)}),$$

and using an expansion similar as in (7.18) we obtain with $n' = n - |w|$,

$$nVar(\hat{\xi}|w) = \sum_{x_1,x_2 \in \mathbf{X}} x_1 x_2 \frac{1}{P_{c_0}^2(w)} nCov(n'^{-1}N(x_1 w), n'^{-1}N(x_2 w)) + o(1).$$

Similar to (7.20) we then get with $m = |w|$,

$$\begin{aligned} nVar(\hat{\xi}|w) &= \frac{1}{P_{c_0}^2(w)} \sum_{x_1,x_2 \in \mathbf{X}} x_1 x_2 \sum_{k=-\infty}^{\infty} Cov(1_{[X_0^m = x_1 w]}, 1_{[X_k^{k+m} = x_2 w]}) + o(1) \\ &= \frac{1}{P_{c_0}(w)} \sum_{x_1,x_2 \in \mathbf{X}} x_1 x_2 P_{c_0}(x_2|w) \sum_{k=-\infty}^{\infty} (\mathbb{P}_{P_{c_0}}[X_0^m = x_1 w | X_k^{k+m} = x_2 w] - P_{c_0}(x_1 w)) \\ &+ o(1). \end{aligned}$$

Thus, by integrating over $w = c(Y_{-\infty}^n)$, $nI_n = C(\tau_c, P_{c_0}) + o(1)$. This, together with (7.17) and (7.22) completes the proof. □

Proof of Theorem 3.2. The decomposition $\mathrm{FPE}_\delta(\tau_c) = S + B + V_n$ follows by the definitions. It remains to analyze the V_n term. We write

$$\begin{aligned} |V_n| &= |\mathbb{E}[1_{[Y_{n+1} \neq \mathrm{AM}_{P_{c_0}}(c(Y_{-\infty}^n))]} - 1_{[Y_{n+1} \neq \varphi(c(Y_{-\infty}^n), X_1^n)]} | c(Y_{-\infty}^n)]| \\ &\leq \mathbb{E}[1_{[\varphi(c(Y_{-\infty}^n), X_1^n) \neq \mathrm{AM}_{P_{c_0}}(c(Y_{-\infty}^n))]} | c(Y_{-\infty}^n)]. \end{aligned} \tag{7.23}$$

We now fix $w = c(Y_{-\infty}^n)$. By assumption (B1),

$$\mathbb{P}[\varphi(w, X_1^n) \neq \mathrm{AM}_{P_{c_0}}(w)|w] \leq \mathbb{P}[\max_{x \in \mathbf{X}} |\hat{P}_c(x|w) - P_{c_0}(x|w)| > \varepsilon/2|w]. \tag{7.24}$$

Similarly as in (7.18) we get with $n' = n - |w|$,

$$\begin{aligned} & \hat{P}_c(x|w) - P_{c_0}(x|w) \\ = & \frac{1}{P_{c_0}(w)}(n'^{-1}N(xw) - P_{c_0}(xw)) - \frac{n'^{-1}N(xw)}{\tilde{P}^2(w)}(n'^{-1}N(w) - P_{c_0}(w)) \\ = & I_n - II_n, \end{aligned} \tag{7.25}$$

where $\tilde{P}(w) = P_{c_0}(w) + \nu(n'^{-1}N(w) - P_{c_0}(w)),\ 0 < \nu < 1$.
Consider the sets

$$\begin{aligned} D_n(x, w) &= \{|n'^{-1}N(xw) - P_{c_0}(xw)| > P_{c_0}(xw)\varepsilon/6\} \\ E_n(w) &= \{|n'^{-1}N(w) - P_{c_0}(w)| > P_{c_0}(w)\varepsilon/6\}. \end{aligned}$$

Then,

$$|I_n| \leq \varepsilon/6 P_{c_0}(x|w) \leq \varepsilon/6 \text{ on } D_n^C(x, w). \tag{7.26}$$

For the second term II_n in (7.25), consider first $\frac{n'^{-1}N(xw)}{\tilde{P}^2(w)}$. The denominator can be bounded on $E_n^C(w)$ as

$$\tilde{P}^2(w) \geq P_{c_0}(w)^2(1 - \varepsilon/6)^2 \geq P_{c_0}(w)^2 25/36,$$

since $\varepsilon \leq 1$.
For the numerator, on $E_n^C(w)$,

$$n'^{-1}N(xw) \leq P_{c_0}(w)(1 + \varepsilon/6) \leq P_{c_0}(w)7/6,$$

since $\varepsilon \leq 1$.
Thus, on $D_n^C(x, w) \cap E_n^C(w)$,

$$\frac{n'^{-1}N(xw)}{\tilde{P}^2(w)} \leq \frac{2}{P_{c_0}(w)}.$$

On the other hand, on $E_n^C(w)$, $|n'^{-1}N(w) - P_{c_0}(w)| \leq P_{c_0}(w)\varepsilon/6$. Thus,

$$|II_n| \leq \varepsilon/3 \text{ on } D_n^C(x, w) \cap E_n^C(w). \tag{7.27}$$

Therefore, by (7.25)-(7.27),

$$\max_{x\in\mathbf{X}} |\hat{P}_c(x|w) - P_{c_0}(x|w)| > \varepsilon/2 \text{ on } \{\cup_{x\in\mathbf{X}} D_n(x,w)\} \cup E_n(w),$$

and thus

$$\mathbb{P}[\max_{x\in\mathbf{X}} |\hat{P}_c(x|w) - P_{c_0}(x|w)| > \varepsilon/2|w] \leq \sum_{x\in\mathbf{X}} \mathbb{P}[D_n(x,w)] + \mathbb{P}[E_n(w)]. \tag{7.28}$$

By formula (7.23), (7.24) and (7.28),

$$|V_n| \leq |\mathbf{X}| \max_{x\in\mathbf{X}, w\in\tau_c} \mathbb{P}[D_n(x,w)] + \max_{w\in\tau_c} \mathbb{P}[E_n(w)]. \tag{7.29}$$

It remains to give some uniform bounds for $\mathbb{P}[D_n(x,w)]$ and $\mathbb{P}[E_n(w)]$. For the set $D_n(x,w)$, we write

$$|n'^{-1}N(xw) - P_{c_0}(xw)| \leq |n'^{-1}N(xw) - \mathbb{E}[n'^{-1}N(xw)]| + P_{c_0}(xw)/n'.$$

Thus, for $n' > 30/\varepsilon$, $P_{c_0}(xw)/n' < \varepsilon/30P_{c_0}(xw)$. Hence for $n' > 30/\varepsilon$, $|n'^{-1}N(xw) - \mathbb{E}[n'^{-1}N(xw)]| > P_{c_0}(xw)\varepsilon/5$ implies $|n'^{-1}N(xw) - P_{c_0}(xw)| > P_{c_0}(xw)\varepsilon/6$. We then consider the sets

$$\tilde{D}_n(x,w) = \{|n'^{-1}N(xw) - \mathbb{E}[n'^{-1}N(xw)]| > P_{c_0}(xw)\varepsilon/5\} \supseteq D_n(x,w) \text{ for } n' > 30/\varepsilon.$$

Now, we employ some exponential inequalities to bound the probabilities for $E_n(w)$ and $\tilde{D}_n(x,w)$. We follow a technique described in Doukhan (1994, Proposition 2, Ch. 1.4.2). For both type of sets we use $\sigma = A(\log(n-k_0+1))^{1/2}$ (A a constant) in the notation of Doukhan. Note that assumption (A1) implies for the ϕ-mixing coefficients $\phi_{P_{c_0}}(k) \leq (1-\kappa)^k$. Thus, in the notation of Doukhan's Proposition 2, $k_n \leq C(\kappa)\log(n-k_0+1)$, $C(\kappa) > 0$ a constant depending on κ. For the sets $E_n(w)$ and $\tilde{D}_n(x,w)$ we have in Doukhan's notation $x = P_{c_0}(w)\varepsilon\sqrt{n'}/(6\sigma)$ and $x = P_{c_0}(xw)\varepsilon\sqrt{n'}/(5\sigma)$, respectively. Then, for A sufficiently large, the restriction in Doukhan $0 \leq x \leq \frac{\sigma\sqrt{n'}}{8bk_n}$ holds. Note that $n' \geq n-k_0+1$ and by assumption (B1), $P_{c_0}(xw) \geq \pi$ for all xw. Then, Proposition 2 in Doukhan (1994, Ch. 1.4.2) yields for $n' > 30/\varepsilon$, i.e., for n sufficiently large,

$$\max_{x\in\mathbf{X}, w\in\tau_c} \mathbb{P}[\tilde{D}_{n,w}] \leq C_1 \exp(-C_2(\kappa)\varepsilon^2\pi^2(n-k_0+1)/\log(n-k_0+1)),$$

and an even better bound (a bigger constant $C_2(\kappa)$) applies for $\max_{w\in\tau_c} \mathbb{P}[E_n(w)]$. These bounds, together with (7.29) complete the proof. □

Proof of Theorem 3.3. We decompose

$$\mathrm{KLI}(\tau_c)/n = B_n + V_n/n,\ V_n = \int_{\mathbf{X}^n} \log\left(\frac{\bar{P}_c(y_1^n)}{\hat{P}_c(y_1^n)}\right) dP_{c_0}(y_1^n). \tag{7.30}$$

It is then helpful to parameterize the probability measures on $\mathbf{X}^\infty$ as $\bar{P}_c = P_{(c,\bar{\theta})}$, $\hat{P}_c = P_{(c,\hat{\theta})}$, $P_{c_0} = P_{(c_0,\theta_0)}$, where $\bar{\theta}$, $\hat{\theta}$ and θ_0 are the transitions probabilities on τ_c and τ_{c_0},

respectively. Without loss of generality we assume $\mathbf{X} = \{0, \ldots, |\mathbf{X}| - 1\}$: then, these transition probabilities are indexed as

$$
\begin{aligned}
&(\bar{\theta})_{wx} = P_{c_0}(x|w) = P_{c_0}(xw)/P_{c_0}(w), \;\; w \in \tau_c, \\
&(\hat{\theta})_{wx} = \hat{P}_c(x|w) = N(xw)/N(w), \;\; w \in \tau_c \text{ (the MLE on } \tau_c), \\
&(\theta_0)_{wx} = P_{c_0}(x|w) = P_{c_0}(xw)/P_{c_0}(w), \;\; w \in \tau_{c_0}.
\end{aligned}
$$

As in standard maximum likelihood theory we develop

$$
\begin{aligned}
&\log(P_{(c,\hat{\theta})}(y_1^n)) = \log(P_{(c,\bar{\theta})}(y_1^n)) + U_{(c,\bar{\theta})}(y_1^n)^T(\hat{\theta} - \bar{\theta}) + 1/2(\hat{\theta} - \bar{\theta})^T H_{(c,\tilde{\theta})}(y_1^n)(\hat{\theta} - \bar{\theta}), \\
&\|\tilde{\theta} - \bar{\theta}\| \leq \|\hat{\theta} - \bar{\theta}\|,
\end{aligned}
$$

where $U_{(c,\bar{\theta})}(y_1^n) = \frac{\partial}{\partial\theta} \log(P_{(c,\theta)}(y_1^n))|_{\theta=\bar{\theta}}$ is the score statistic at $\bar{\theta}$ and $H_{(c,\tilde{\theta})}(y_1^n) = \frac{\partial^2}{\partial\theta\partial\theta^T} \log(P_{(c,\theta)}(y_1^n))|_{\theta=\tilde{\theta}}$ is the Hessian matrix at $\tilde{\theta}$.
Since $\mathbb{E}[U_{(c,\bar{\theta})}(Y_1^n)] = \int_{\mathbf{X}^n} U_{(c,\bar{\theta})}(y_1^n) dP_{(c_0,\theta_0)}(y_1^n) = 0$ we have by (7.30),

$$
V_n = -1/2(\hat{\theta} - \bar{\theta})^T \int_{\mathbf{X}^n} H_{(c,\tilde{\theta})}(y_1^n) dP_{(c_0,\theta_0)}(\hat{\theta} - \bar{\theta}). \tag{7.31}
$$

For the MLE $\hat{\theta}$ we consider first the score statistic

$$
\begin{aligned}
U_{(c,\theta)}(X_1^n) &= \sum_{t=k_0+1}^{n} \tilde{U}_{(c,\theta)}(X_{t-k_0}^t) + o_P(1), \\
\tilde{U}_{(c,\theta)}(X_{t-k_0}^t) &= \frac{\partial}{\partial\theta} \log(P_{(c,\theta)}(X_t|c(X_{t-k_0}^{t-1})) = \frac{\partial}{\partial\theta} \log(\theta)_{c(X_{t-k_0}^{t-1}),X_t}.
\end{aligned}
$$

At $\bar{\theta}$ and for the component index wx,

$$
\left(\tilde{U}_{(c,\bar{\theta})}(x_{t-k_0}^t)\right)_{wx} = \frac{1}{\bar{\theta}_{wx}} 1_{[x_t=x, c(x_{t-k_0}^{t-1})=w]} - \frac{1}{1 - \sum_{r=0}^{|\mathbf{X}|-1} \bar{\theta}_{wr}} 1_{[x_t=|\mathbf{X}|-1, c(x_{t-k_0}^{t-1})=w]}.
$$

It follows that $\mathbb{E}_{P_{(c_0,\theta_0)}}[\tilde{U}_{(c,\bar{\theta})}(X_{t-k_0}^t)] = 0$. Then, by the geometric mixing property of P_{c_0} (see also remark 7.1),

$$
\begin{aligned}
n^{-1/2} \sum_{t=k_0+1}^{n} \tilde{U}_{(c,\bar{\theta})}(X_{t-k_0}^t) &\Rightarrow \mathcal{N}(0, F(c,\bar{\theta})), \\
F(c,\bar{\theta}) &= \sum_{m=-\infty}^{\infty} \mathbb{E}[\tilde{U}_{(c,\bar{\theta})}(X_{-k_0}^0) \tilde{U}_{(c,\bar{\theta})}^T(X_{m-k_0}^m)].
\end{aligned} \tag{7.32}
$$

Note that if $\tau_c = \tau_{c_0}$, that is under the true model, then $\bar{\theta} = \theta_0$ and we can exploit the Markov structure so that $F(c,\bar{\theta}) = \mathbb{E}[\tilde{U}_{(c,\bar{\theta})}(X_{-k_0}^0) \tilde{U}_{(c,\bar{\theta})}(X_{k_0}^0)^T]$.
The Hessian matrix in (7.31) is of the form

$$
\begin{aligned}
&\left(H_{(c,\theta)}(y_1^n)\right)_{w_1x_1, w_2x_2} \\
= \;& -\delta_{w_1w_2} \sum_{t=k_0+1}^{n} (\delta_{x_1x_2} \frac{1}{\theta_{w_1x_1}^2} 1_{[y_t=x_1, c(y_{t-k_0}^{t-1})=w_1]} + \frac{1}{(1 - \sum_{r=0}^{|\mathbf{X}|-1} \theta_{w_1r})^2} 1_{[y_t=|\mathbf{X}|-1, c(y_{t-k_0}^{t-1})=w_1]} \\
+ \;& o(1).
\end{aligned}
$$

Thus, the limit of the expected value is given by

$$\begin{aligned} J(c,\bar{\theta}) &= \lim_{n\to\infty} n^{-1} \int_{\mathbf{X}^n} H_{(c,\bar{\theta})}(y_1^n) dP_{(c_0,\theta_0)}(y_1^n) \\ &= -\delta_{w_1 w_2}(\delta_{x_1 x_2} \frac{1}{\bar{\theta}_{w_1 x_1}} + \frac{1}{1-\sum_{r=0}^{|\mathbf{X}|-1} \bar{\theta}_{w_1 r}}) P_{(c_0,\theta_0)}(w_1). \end{aligned} \tag{7.33}$$

It is straightforward to show $\tilde{\theta} = \bar{\theta} + o_P(1)$. We then get for the expression in (7.31),

$$\int_{\mathbf{X}^n} n^{-1} H_{(c,\tilde{\theta})}(y_1^n) dP_{(c_0,\theta_0)}(y_1^n) = J(c,\bar{\theta}) + o_P(1). \tag{7.34}$$

Also, by standard arguments for MLE, using (7.32), (7.33) and the mixing property of P_{c_0}, we get

$$n^{1/2}(\hat{\theta} - \bar{\theta}) \Rightarrow -J(c,\bar{\theta})^{-1} F(c,\bar{\theta})^{1/2} Z, \;\; Z \sim \mathcal{N}_{D(\tau_c)}(0, I). \tag{7.35}$$

Thus, by (7.31), (7.34) and (7.35) we get

$$V_n \Rightarrow 1/2 Z^T F(c,\bar{\theta})^{1/2} J(c,\bar{\theta})^{-1} F(c,\bar{\theta})^{1/2} Z.$$

Since $\bar{\theta}$ is a function of P_{c_0} on τ_c, and since the quantities $F(.,.)$ in (7.32) and $J(.,.)$ in (7.33) implicitly also depend on P_{c_0} we set $\Sigma(\tau_c, P_{c_0}) = F(c,\bar{\theta})^{1/2} J(c,\bar{\theta})^{-1} F(c,\bar{\theta})^{1/2}$. This, together with (7.30) completes the proof. □

Note that if $\tau_c = \tau_{c_0}$, then $\bar{\theta} = \theta_0$ and $F(c_0,\theta_0) = J(c_0,\theta_0)$. Then, $\Sigma(\tau_{c_0}, P_{c_0}) = I_{D(\tau_{c_0})}$ and $V_n \Rightarrow 1/2\chi^2_{D(\tau_{c_0})}$.

For proving the Theorems in section 4, we first restate a result about the context algorithm in section 2.1.

Lemma 7.1 *Consider a finite realization X_1^n from P_{c_0}, satisfying (A1) and (A2). Assume that the cut-off $K_n > (2|\mathbf{X}|+3)\log(n)$ in Step 2 of the context algorithm for constructing the estimate $\hat{P}_{\hat{c}_0}$ in (2.6). Then,*

(i) $\mathbb{P}_{P_{c_0}}[\hat{c}_0(.) = c_0(.)] = 1 + o(n^{-1})$ $(n \to \infty)$,

(ii) $\hat{P}_{\hat{c}_0}(x_1^m) = P_{c_0}(x_1^m) + o_P(1)$ *for all* $x_1^m \in \mathbf{X}^m$ $(m \in \mathbb{N})$,

(iii) On a set A_n with $\mathbb{P}_{P_{c_0}}[A_n] \to 1$ $(n \to \infty)$, $\hat{P}_{\hat{c}_0}$ satisfies (A1) with κ replaced by $\kappa/2$ and (A2).

Proof: The assertions (i) and (ii) are special cases of Theorems 3.1, 5.1 and 5.2 in Bühlmann and Wyner (1997). Assertion (iii) follows from formula (5.16) in Bühlmann and Wyner (1997). □

Remark 7.1. Assertion (iii) of Lemma 7.1 implies the geometric ϕ-mixing property of $\hat{P}_{\hat{c}_0}$ with $\phi_{\hat{P}_{\hat{c}_0}}(k) \leq (1-\kappa/2)^k$ on the set A_n.

Proof of Theorem 4.1. By Lemma 7.1, the bootstrapped process $(X_t^*)_{t\in\mathbb{Z}} \sim \hat{P}_{\hat{c}_0}$ satisfies again (A1) and (A2) on a set A_n with $\mathbb{P}[A_n] \to 1$. Therefore, by the same arguments as in the proof of Theorem 3.1, the decomposition $\mathrm{FPE}^*_{L_2}(\tau_c) = S^* + B^* + V_n^*$ holds on

the set A_n. It remains to show the convergence of S^*, B^*, V_n^* to S, B and $C(P_{c_0}, \tau_c)$, respectively.

The convergences $S^* = S + o_P(1)$ and $B^* = B + o_P(1)$ follow directly by the finiteness of τ_c, τ_{c_0} and Lemma 7.1(ii).
By using Lemma 7.1(iii) we get as for analyzing nV_n in the proof of Theorem 3.1,

$$nV_n^* = C(\tau_c, \hat{P}_{\hat{c}_0}) + o_P(1) \ (n \to \infty).$$

Using the geometric ϕ-mixing property of $\hat{P}_{\hat{c}_0}$ on the set A_n (see remark 7.1) we obtain $C(\tau_c, \hat{P}_{\hat{c}_0}) = C(\tau_c, P_{c_0}) + o_P(1)$, which then implies $nV_n^* = nV_n + o_P(1)$. □

Proof of Theorem 4.2. As in the proof of Theorem 4.1 we rely again on Lemma 7.1. The decomposition $\mathrm{FPE}_\delta^*(\tau_c) = S^* + B^* + V_n^*$ follows by the definitions.

By Lemma 7.1(i) and (ii) and the finiteness of τ_c and τ_{c_0} we obtain the convergences $S^* = S + o_P(1)$ and $B^* = B + o_P(1)$.
Again by Lemma 7.1(i) and (ii) assumption (B1) with P_{c_0} replaced by $\hat{P}_{\hat{c}_0}$ holds in probability. Finally by using Lemma 7.1(iii), which implies the geometric ϕ-mixing property for $\hat{P}_{\hat{c}_0}$ on the set A_n (see remark 7.1), we get the exponential bound in probability, as for analyzing V_n in the proof of Theorem 3.2. □

Proof of Theorem 4.3. The decomposition $\mathrm{KLI}^*(\tau_c)/n = B_n^* + V_n^*/n$ is immediate. The convergence $B_n^* = B_n + o_P(1)$ follows by Lemma 7.1(i)-(ii) and the finiteness of τ_{c_0} and τ_c.
It remains to show the proper convergence for V_n^*. By Lemma 7.1(iii) we can carry out the same steps as in the proof of Theorem 3.3 to obtain

$$\begin{aligned} &\mathbb{P}_{\hat{P}_{\hat{c}_0}}[V_n^* \le x] = \mathbb{P}[1/2 Z^T \Sigma(\tau_c, \hat{P}_{\hat{c}_0}) Z \le x | \hat{P}_{\hat{c}_0}] + o_P(1), \ x \in \mathbb{R}, \\ &\Sigma(\tau_c, \hat{P}_{\hat{c}_0}) = F(c, \bar{\theta}^*)^{1/2} J(c, \bar{\theta}^*)^{-1} F(c, \bar{\theta}^*)^{1/2}, \end{aligned} \tag{7.36}$$

with $F(.,.)$ as in (7.32) and $J(.,.)$ as in (7.33), but with $P_{(\hat{c}_0, \hat{\theta}_0)}$ instead of $P_{(c_0, \theta_0)}$: here $(\bar{\theta}^*)_{wx} = \hat{P}_{\hat{c}_0}(wx)/\hat{P}_{\hat{c}_0}(w)$, $w \in \tau_c$.
By Lemma 7.1(i)-(ii) we then get

$$\begin{aligned} F(c, \bar{\theta}^*) &= F(c, \bar{\theta}) + o_P(1), \\ J(c, \bar{\theta}^*) &= J(c, \bar{\theta}) + o_P(1), \end{aligned}$$

and thus $\Sigma(\tau_c, \hat{P}_{\hat{c}_0}) = \Sigma(\tau_c, P_{c_0}) + o_P(1)$. Together with (7.36), this completes the proof.□

Acknowledgments: I thank Richard Olshen for a helpful discussion about pruning in tree structured models and Adi Wyner for many general conversations about variable length Markov chains.

References

[1] Akaike, H. (1969). Fitting autoregressive models for prediction. Ann. Inst. Statist. Math. **21** 243-247.

[2] Akaike, H. (1970). Statistical predictor identification. Ann. Inst. Statist. Math. **22** 202-217.

[3] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone. C.J. (1984). Classification and Regression Trees. Wadsworth.

[4] Bühlmann, P. (1997). Efficiency and adaptivity of the context algorithm for variable length Markov chains. Preprint, Seminar für Statistik, ETH Zürich, Switzerland.

[5] Bühlmann, P. and Wyner, A.J. (1997). Variable length Markov chains. Tech Rep. 479, Dept. Statist., University of California, Berkeley.

[6] Cavanaugh, J. and Shumway, R. (1997). A bootstrap variant of AIC for state-space model selection. Statistica Sinica **7** 473-496.

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

[8] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. J. Amer. Statist. Assoc. **78** 316-331.

[9] Efron, B. (1986). How biased is the apparent error rate of a prediction rule. J. Amer. Statist. Assoc. **81** 461-470.

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

[11] Rissanen, J.J. (1994). Noise separation and MDL modeling of chaotic processes. In From Statistical Physics to Statistical Inference and Back (Eds. P. Grassberger and J.-P. Nadal), pp. 317-330. Kluwer.

[12] Shibata, R. (1989). Statistical aspect of model selection. In From Data to Model (Ed. J.C. Willems), pp. 215-240. Springer.

[13] Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. Statistica Sinica **7** 375-394.

[14] Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. Suri-Kagaku (Mathematical Sciences) **153** 12-18. In Japanese.

[15] Tong, H. (1975). Determination of the order of a Markov chain by Akaike's information criterion. J. Appl. Prob. **12** 488-497.

[16] Weinberger, M.J., Rissanen, J.J. and Feder, M. (1995). A universal finite memory source. IEEE Trans. Inform. Theory **IT-41** 643-652.

[17] Weinberger, M.J., Rissanen, J.J. and Arps, R.B. (1996). Applications of universal context modeling to lossless compression of gray-scale images. IEEE Trans. Image Proc. **IP-5** 575-586.

Seminar für Statistik, SOL
ETH Zentrum
CH-8092 Zürich
Switzerland
e-mail: buhlmann@stat.math.ethz.ch