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

published in *Annals of the Institute of Statistical Mathematics*, 2000, Vol. 52, pp. 287–315.

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

January 1999

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 criteria. In particular, we obtain a data-driven tuning of Rissanen's tree structured context algorithm which is a computationally feasible procedure for selection and estimation of a VLMC.

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

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 $\mathcal{X}$. The processes in this class are still Markovian of higher order, but their memory can have variable length. They are also known under the names 'tree models', 'FSMX models' or 'finite-memory sources', cf. Rissanen (1986), Weinberger et al. (1992) and Weinberger et al. (1995). 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 when estimating the unknown transition mechanism. 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 asymptotically efficient in the sense of coding, cf. Weinberger et al. (1995), and also in a more statistical sense for estimating a smooth functional, cf. Bühlmann (1998). Successful applications of the context algorithm have been reported among others by Rissanen (1994) for modeling chaotic processes and by Weinberger et al. (1996) and Bunton (1997) for data compression.

This paper addresses two further problems which are closely connected to each other and which play an eminent role when fitting a VLMC with a finite amount of data. We sometimes refer to them as the model selection problem for VLMC's.

Problem 1. How can we generally measure in a data-driven way model complexity in the class of VLMC's? The word 'general' refers here to various aims for which an estimated model is used: they can be formalized in terms of various risk functions.

Problem 2. What is a computationally feasible way to estimate a (sub-)optimal, with respect to a risk function as mentioned above, member in the combinatorially very large class of VLMC's? In view of Problem 1, this question involves some (restricted) minimization of estimated model complexity measures.

Regarding Problem 1, the following should motivate a more rigorous analysis than asymptotic consistency of model or state selection in the class of VLMC's. For finite sample size, the true structure of a model, here the true minimal state space of a VLMC, is not necessarily optimal in terms of minimizing a risk for estimating the whole probability distribution of the true underlying VLMC (or a functional thereof). Thus, even under knowledge of the true model structure we may not want to use it for estimating the true underlying process, and consistency for the model structure (which we then hypothetically would have by knowing the true structure) is not always relevant. The phenomenon corresponds to a bias-variance trade-off, accounting for additional variance when estimating additional parameters in larger models which have smaller bias. Many commonly known model selection techniques are based on a goodness of fit measure and a penalty term, the latter taking the high variance in large models into account. But the context algorithm is not of this nature: it makes local test decisions which can be proven to be consistent for estimation of the underlying VLMC. However, this local decision approach then never takes a global view aiming to minimize an overall risk (for finite sample sizes), for example with a penalized goodness of fit measure. A solution to Problem 1 in the case of finite-state (FS) models has been considered by using the minimum description length (MDL) criterion in Weinberger and Feder (1994). They show that the final estimate for the whole probability distribution of the underlying process, based on the estimated model structure via MDL, achieves an asymptotic lower bound in terms of per-symbol code-length (Rissanen, 1986). However, minimizing an MDL criterion over the class of FS models is computationally much too complex to be ever realized, which relates to Problem 2, and the above result is mainly of theoretical interest. We study here model complexity, and also selection of a VLMC, by measuring statistical performance of the estimated distribution of a VLMC with a general risk function, such as final prediction error with the quadratic or the zero-one loss or the expected Kullback-Leibler information. By specifying a certain risk function we can tailor the model selection problem towards specific aims. The estimation of the various risks, and thus of model complexity, can be done consistently with a resampling scheme.

Regarding Problem 2, we propose a tuning of the non-predictive context algorithm with respect to some risk function, or measure of model complexity, as mentioned in connection with Problem 1. When using the context algorithm 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. Similar to estimation of a risk function, we propose a resampling technique for estimating a cut-off which aims for global optimality of a VLMC model, in contrast to only considering local test decisions which are the basis of the context algorithm. The optimality of the cut-off is with respect to a chosen risk function, as in Problem 1. Searching for an optimal cut-off is a computationally feasible task: varying over a real-valued cut-off parameter produces finitely, and not extremely many VLMC tree models. This operation achieves a similar task as 'cost-complexity pruning' in CART (Breiman et al., 1984, Ch. 3.3). Our approach thus equips the intrinsic local nature of the context algorithm based on test decisions with a global optimality criterion: the local nature of the algorithm is crucial for obtaining a computationally feasible procedure, the global optimality view is crucial for obtaining at least a suboptimal solution to the computationally intractable problem of minimizing a model complexity criterion among all VLMC submodels, say of dimension less than a reasonable bound of smaller order than the sample size. Our approach yields then a practical data-driven rule for determining the cut-off tuning parameter in the non-predictive context algorithm.

A very interesting alternative proposal uses the context algorithm in a predictive way, cf. Rissanen (1994), Weinberger et al. (1996) and Bunton (1997). This predictive scheme does not require selection of a cut-off tuning parameter as in the non-predictive case. Model complexity for VLMC's is now estimated by predictive losses which provides an answer to Problem 1. A remarkable answer to Problem 2 with the predictive context algorithm is described in Bunton (1997): by using dynamic programming, the loss is computed in every predictive step on some global basis. This can be implemented in conjunction with a general predictive loss function and thus shares the same wide applicability as our approach. Asymptotic properties of such a predictive scheme are unknown so far. We discuss some open questions in the last paragraph of section 6. In the sequel, we focus on the non-predictive case: it fits into the framework of model selection with classical maximum likelihood estimation for unknown parameters which is a very common set-up in non-sequential applications.

In the combinatorially simple case of estimating the order of a full Markov chain, Tong (1975) has proposed the Akaike information criterion (AIC) which should aim to minimize an expected Kullback-Leibler information. This proposal is improved by our general resampling strategy in the special case of order selection in classical full Markov chains. The problem of order selection for full Markov chains has also been considered by Merhav et al. (1989). They don't consider a risk for estimating the true underlying Markov chain (or a functional thereof) but rather a minimization of the underestimation probability of the true order constrained to an upper bound for the probability of an overestimation event. This approach is thus mainly concerned with finding the true model structure but not very much with statistical performance (in terms of a risk as mentioned in Problem 1) when using the estimated distribution of the Markov chain.

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 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 define now 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 $\mathcal{X}$. Thus,

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

Such full Markov chains are very hard to estimate since they involve $|\mathcal{X}|^k(|\mathcal{X}|-1)$ free parameters, where $|\mathcal{X}|$ denotes the cardinality of $\mathcal{X}$. To get less complex models, the idea is to lump irrelevant states in the history X^0_{-k+1} 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 $X^{t-1}_{-\infty}$ 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 $X^{t-1}_{-\infty}$. 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 $X_t \in \mathcal{X}$, $|\mathcal{X}| < \infty$. Denote by $c : \mathcal{X}^\infty \to \cup_{m=0}^\infty \mathcal{X}^m$ a (variable projection) function,*

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

Then, $c(.)$ is called a context function and for any $t \in \mathbb{Z}$, $c(x^{t-1}_{-\infty})$ 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 $X_t \in \mathcal{X}$, $|\mathcal{X}| < \infty$ and corresponding context function $c(.)$ as given in Definition 2.1. Let $0 \le k \le \infty$ be the smallest integer such that*

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

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 $\mathcal{X}^\infty$.

Instead of the name VLMC, the terminology tree model, FSMX model or finite-memory source has also been used, cf. Weinberger et al. (1992) and Weinberger et al. (1995). 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,

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

In retrospect, we could define a context function $c(.) : \mathcal{X}^k \to \cup_{m=0}^k \mathcal{X}^m$, 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 $\mathcal{X}^\infty$ and sometimes on $\mathcal{X}^k$. The context function projects the k-th (or infinite) order history x_{-k+1}^0 into $\cup_{m=0}^k \mathcal{X}^m$. Often the range space of the context function $c(.)$ is not the full space $\mathcal{X}^k$, 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 $|\mathcal{X}|$ offsprings. Then, each value of a context function $c(.) : \mathcal{X}^k \to \cup_{m=0}^k \mathcal{X}^m$ can be represented as a branch 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 $x_{-\ell(x_0,\ldots,x_{-k+1})+1}$. Note that context trees do not have to be complete, i.e., every internal node does not need to have exactly $|\mathcal{X}|$ offsprings.

Example 2.1 *$|\mathcal{X}| = 2$, $k = 3$.*

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 *in Figure 2.1 A 'growing to the left' sub-branch represents the symbol* 0 *and vice versa for the symbol* 1.

Figure 2.1 about here

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

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

The context tree τ_c is nothing else than the minimal state space of the VLMC P_c (we sometimes refer to the elements of τ_c as branches and sometimes as nodes in a tree). 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, and states $w \in \tau_c$ do not need to correspond to terminal nodes in τ_c. But we can reconstruct the context function $c(.)$ from either τ_c or τ^t_c. Note that an internal node with $b < |\mathcal{X}|$ offsprings can be implicitly thought to be complete by adding one complementary offspring, lumping the $|\mathcal{X}| - b$ 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 denote by wu the concatenation of two strings $w, u \in \cup_{m=1}^{\infty} \mathcal{X}^m$, written in reverse time $wu = (\ldots, w_2, w_1, \ldots, u_2, u_1)$; also, we always make the convention that quantities involving time indices $t \notin \{1, \ldots, n\}$ equal zero (or are irrelevant). Let

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

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

$$\hat{P}(w) = N(w)/n,\ \hat{P}(v|w) = \frac{N(vw)}{N(w)},\ v, w \in \cup_{m=1}^{\infty} \mathcal{X}^m. \tag{2.3}$$

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

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

with $K = K_n \to \infty$ $(n \to \infty)$ a cut-off to be chosen by the user. The idea behind this strategy is to search for the largest state space such that its terminal nodes wu (in the tree representation) have sufficiently different transition probabilities, compared to their parent nodes w and measured with a scaled Kullback-Leibler information between $\hat{P}(.|wu)$ and $\hat{P}(.|w)$.

Step 1 Given data $X_1, \ldots, X_n$ taking values in a finite space $\mathcal{X}$, fit a maximal ($|\mathcal{X}|$-ary) context tree, i.e., search for the context function $c_{max}(.)$ with terminal node context tree representation τ^t_{max}, where τ^t_{max} is the biggest tree such that every element (terminal node) in τ^t_{max} has been observed at least twice in the data. This can be formalized as follows:
τ^t_{max} is such that $w \in \tau^t_{max}$ implies $N(w) \geq 2$, and such that for every τ^t, where $w \in \tau^t$ implies $N(w) \geq 2$, it holds that $\tau^t \preceq \tau^t_{max}$.
($\tau^t_1 \preceq \tau^t_2$ means: $w \in \tau^t_1 \Rightarrow wu \in \tau^t_2$ for some $u \in \cup_{m=0}^{\infty} \mathcal{X}^m$ ($\mathcal{X}^0 = \emptyset$)).
Set $\tau^t_{(0)} = \tau^t_{max}$.

Step 2 Examine every element (terminal node) of $\tau^t_{(0)}$ as follows (the order of examining is irrelevant). Let $c(.)$ be the corresponding context function 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 $w = \emptyset$, i.e., the root node).
Replace the context $wu = x^0_{-\ell+1}$ by $w = x^0_{-\ell+2}$ if

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

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

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

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 $P_c(x_1|c(x^0_{-\infty})) = \mathbb{P}_{P_c}[X_1 = x_1|c(X^0_{-\infty}) = c(x^0_{-\infty})]$ by $\hat{P}(x_1|\hat{c}(x^0_{-\infty}))$, where $\hat{P}(.|.)$ is defined as in (2.2).

Step 1 in the context algorithm ensures a large tree τ^t_{max} as a basis to start the pruning process in Step 2. The construction of τ^t_{max} is fast and simple, the requirement $N(w) \geq 2$ for all $w \in \tau^t_{max}$ guarantees at least two observations per branch w (the lower bound 2 for $N(w)$ accepts any potentially interesting branch w whose relevance from the data is supported by at least two observations). The pruning in Step 2 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} goes back to Weinberger et al. (1995). We denote by

(2.5) $\hat{P}_c$ the maximum likelihood (ML) fitted VLMC, given $c(.)$ or τ_c,

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

Note that the ML fitted VLMC 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 a setting where the dimension of the true underlying process is allowed to grow with increasing sample size 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 in the statistical sense is given in Bühlmann (1998). 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 $K_n \sim C\log(n),\ C > (2|\mathcal{X}| + 4)$. In this paper we also develop a strategy for estimating this cut-off K_n as the minimizer of certain risk functions.

The context algorithm as given above is defined on the whole available data sequence X_1^n in a non-predictive way. Another version of the context algorithm can be defined in a *predictive* fashion, based on successive observation-strings X_1^i $(i = 1, \ldots, n)$. Such a version driven by a predictive code length difference, which is then related to predictive stochastic complexity, has been employed in Rissanen (1994), Weinberger et al. (1996) and Bunton (1997). For such a predictive scheme, there isn't any need to specify a cut-off parameter as in Step 2 of our version and the problem of cut-off estimation does not appear. The predictive context algorithm could also be optimized with respect to any risk function by considering a predictive risk as a criterion to be minimized. No consistency or optimality result seems to be known for any of these predictive schemes, see also the discussion in section 1. In the following, we focus only on the non-predictive case which fits into the model selection framework with classical maximum likelihood estimation for unknown parameters.

3 Risk functions and candidate 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}. As candidate models we consider VLMC's of finite orders k in the range $0 \le k < \infty$,

$$\mathcal{M} = \{P_c : P_c \text{ a VLMC of order } k,\ 0 \le k < \infty\}.$$

Often in model selection, the relevant feature of a candidate model $P_c \in \mathcal{M}$ is its structure, here given by the context tree τ_c or the context function $c(.)$. We study the problem of model selection in terms of two different risk criteria, the 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

$$\begin{aligned} L_2(Y_{n+1}, \hat{Y}_{n+1}) &= (Y_{n+1} - \hat{Y}_{n+1})^2, \\ \delta(Y_{n+1}, \hat{Y}_{n+1}) &= 1_{[Y_{n+1} \ne \hat{Y}_{n+1}]}. \end{aligned}$$

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. The terminology 'final prediction error' is used inconsistently in the literature. We refer here to FPE as a theoretical quantity, defined below, and not to the alternative use as an estimator. 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 models in $\mathcal{M}$ with context function $c(.)$ is given by

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

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

$$\varphi(c(Y_{-\infty}^n), X_1^n) = \begin{cases} \mathbb{E}_{\hat{P}_c}[Y_{n+1}|c(Y_{-\infty}^n)] \text{ for the } L_2 \text{ loss} \\ \mathrm{AM}_{\hat{P}_c}(c(Y_{-\infty}^n)) \text{ for the } \delta \text{ loss} \end{cases}, \tag{3.1}$$

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 τ_c corresponding to an element in $\mathcal{M}$, τ_{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 $x \in \mathcal{X}$ and all $w \in \tau_c$, cf. Bühlmann and Wyner (1997).
The FPE's for an element $P_c \in \mathcal{M}$ with corresponding context tree τ_c 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 $Y_{-\infty}^{n+1}$ as test set. The following two Theorems describe how the FPE decomposes into an 'oracle part' which is not depending on the model feature 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 $P(x) = \mathbb{P}_P[X_1^m = x]$ $(x \in \mathcal{X}^m)$ and $P(x|w) = P(xw)/P(w)$ $(x, w \in \cup_{m=1}^{\infty} \mathcal{X}^m)$, where P is a probability measure on $\mathcal{X}^{\mathbb{Z}}$. We then make the following assumption.

(A) P_{c_0} satisfies

$$\min_{x \in \mathcal{X}, w \in \tau_{c_0}} P_{c_0}(x|w) > 0.$$

Assumption (A) ensures that the VLMC P_{c_0} is stationary and geometric ϕ-mixing. Further consequences of (A) are given in section 7, formulae (7.1) and (7.2).

Theorem 3.1 *Consider a finite realization X_1^n from P_{c_0} satisfying (A) and with context tree representation τ_{c_0}. Then, for any element of the models in $\mathcal{M}$ with context function c and corresponding tree representation τ_c, 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 (3.1) and

$$nV_n - C(\tau_c, P_{c_0}) = o(1),$$
$$C(\tau_c, P_{c_0}) = \sum_{w \in \tau_c} \sum_{x_1, x_2 \in \mathcal{X}} x_1 x_2 P_{c_0}(x_2|w) \sum_{k=-\infty}^{\infty} \left(\mathbb{P}_{P_{c_0}}[X_{-|w|}^0 = x_1 w | X_{k-|w|}^k = x_2 w] - P_{c_0}(x_1 w)\right).$$

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 (A), implying a Doeblin type condition as given in (7.1), we still can bound the penalty term linearly in $|\tau_c|$ as

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

where $0 < M(\mathcal{X}, k_0) < \infty$ is a constant, depending on the space $\mathcal{X}$ and the order k_0 of the VLMC P_{c_0}.

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

(B) For an element $P_c \in \mathcal{M}$ with corresponding context tree τ_c,

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

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

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

Theorem 3.2 *Consider a finite realization X_1^n from P_{c_0} satisfying (A) and with context tree representation τ_{c_0}. Then, for any element of the models in $\mathcal{M}$ with context function c and corresponding tree representation τ_c, satisfying (B), 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 (3.1) and for n sufficiently large,

$$|V_n| \le |\mathcal{X}| C_1 \exp(-C_2\varepsilon^2\pi^2 \frac{n-k_c}{\log(n-k_c)}),$$

k_c the order of $c(.)$ (the depth of τ_c), $C_1, C_2 > 0$ some constants.

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 $n - k_c$ with k_c the finite order of $c(.)$ (the depth of τ_c) and the size $|\tau_c|$ entering only implicitly into the speed of the exponential decay: larger candidate 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_δ 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)

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

measures a loss between the n-dimensional marginals of P_{c_0} and the maximum likelihood estimate $\hat{P}_c$ of a model $P_c \in \mathcal{M}$ with context tree representation τ_c. 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 KLI(τ_c),

$$\text{EKLI}(\tau_c) = \mathbb{E}_{P_{c_0}}[I_n(P_{c_0}, \hat{P}_c)]. \tag{3.2}$$

Considering the Kullback-Leibler information for model selection has been proposed in the seminal paper of Akaike (1973).

Theorem 3.3 *Consider a finite realization X_1^n from P_{c_0} satisfying (A) and with context tree representation τ_{c_0}. Then, for any element of the models in $\mathcal{M}$ with context function c and corresponding tree representation τ_c, the following decomposition holds:*

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

where $\bar{P}_c$ is induced from P_{c_0} on the model structure τ_c, generated by the transition probabilities

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

$Z \sim \mathcal{N}_{D(\tau_c)}(0, I)$, $D(\tau_c) = |\tau_c|(|\mathcal{X}| - 1)$ the dimension of the candidate model, and $\Sigma(\tau_c, P_{c_0})$ a non-degenerate $D(\tau_c) \times D(\tau_c)$ matrix, depending on the 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 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 the $2nV_n$ 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 nV_n equals $\chi^2_{D(\tau_{c_0})}/2$. The limiting distribution of nV_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 τ_c, being a candidate model in $\mathcal{M}$. 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 model in $\mathcal{M}$ with structure τ_c, compute the bootstrap risk functions,

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

where $\varphi(.,.)$ is as in (3.1) and

$$(Y^*)_1^{n+1} \sim \hat{P}_{\hat{c}_0} \circ \pi^{-1}_{1,\ldots,n+1},$$
$$(X^*)_1^{n+1} \sim \hat{P}_{\hat{c}_0} \circ \pi^{-1}_{1,\ldots,n}, \tag{4.1}$$

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

$$\hat{P}^*_c = T_c((X^*)_1^n) \tag{4.2}$$

is the plug-in version of the ML fitted VLMC $\hat{P}_c = T_c(X_1^n)$ 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 (3.2). In practice, the expectations with respect to $\hat{P}_{\hat{c}_0}$ are evaluated via Monte-Carlo. Minimization of such estimated risks over all models in $\mathcal{M}$ (or all VLMC models having order $0 \le k \le K$ for some K large) with context trees τ_c 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 $K_n > (2|\mathcal{X}|+4)\log(n)$ in Step 2 of the context algorithm for constructing the estimate $\hat{P}_{\hat{c}_0}$. Then,*

$$FPE^*_{L_2}(\tau_c) = S^* + B^* + V^*_n,$$

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

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 $K_n > (2|\mathcal{X}| + 4)\log(n)$ 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 $K_n > (2|\mathcal{X}| + 4)\log(n)$ in Step 2 of the context algorithm for constructing the estimate $\hat{P}_{\hat{c}_0}$. Then,*

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

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 and in case of overestimation. If the numbers of parameters, here given by $D(\tau_c)$, is large, then the V_n-terms are typically not that much negligible compared to the $(S+B)$-terms. For the case of overestimation, let us consider more closely the behavior of the Kullback-Leibler information; the analysis for the FPE is analogous. Assume that

$$\tau_{c_0} \preceq \tau_{c_1} \prec \tau_{c_2}$$

with τ_{c_0} the context tree corresponding to the true underlying P_{c_0}, τ_{c_1} a super-tree of τ_{c_0} (possibly equal to τ_{c_0}) but a sub-tree of τ_{c_2} (for a definition of '$\preceq$' see Step 1 in section 2.1). In this case it can then be easily shown that

$$n^{-1}(\mathrm{KLI}(\tau_{c_2}) - \mathrm{KLI}(\tau_{c_1})) = V_n(\tau_{c_2}) - V_n(\tau_{c_1}) = O_P(n^{-1}), \tag{4.3}$$

where $V_n(\tau_c)$ is the V_n-term in Theorem 3.3 for a context tree τ_c.
For the bootstrapped risks we get under the conditions of Theorem 4.3 an analogous formula,

$$n^{-1}(\mathrm{KLI}^*(\tau_{c_2}) - \mathrm{KLI}^*(\tau_{c_1})) = V_n^*(\tau_{c_2}) - V_n^*(\tau_{c_1}) + o_P(n^{-1})$$

which is of the order n^{-1}. Moreover, by Theorems 3.3 and 4.3 we then obtain,

$$\mathbb{P}^*[(\mathrm{KLI}^*(\tau_{c_2}) - \mathrm{KLI}^*(\tau_{c_1})) \le x] - \mathbb{P}[(\mathrm{KLI}(\tau_{c_2}) - \mathrm{KLI}(\tau_{c_1})) \le x] = o_P(1) \ (x \in \mathbb{R}),$$

establishing the consistency of the bootstrap risk estimator even in the more subtle case where the difference between VLMC models is of the order n^{-1}, see formula (4.3). Such a higher order result is not implied by (and is different from) an optimality result for the context algorithm in Weinberger et al. (1995), considering the per symbol code-length (Rissanen, 1986).

Remark 4.2. The risk $\mathrm{KLI}^*(\tau_c)$ can be related to information criteria such as AIC. Under the assumptions of Theorem 4.3,

$$\begin{aligned} 2\mathrm{EKLI}^*(\tau_c) &= 2\mathbb{E}_{\hat{P}_{\hat{c}_0}}[\mathrm{KLI}^*(\tau_c)|X_1^n] \\ &\approx C - 2\mathbb{E}_{\hat{P}_{\hat{c}_0}}[\log(\hat{P}_c^*((X^*)_1^n))] + 4\mathbb{E}_{\hat{P}_{\hat{c}_0}}[nV_n^*], \end{aligned} \tag{4.4}$$

where $\log(\hat{P}_c^*((X^*)_1^n)$ is the log-likelihood based on $(X^*)_1^n$, i.e., $\hat{P}_c^*(.)$ is estimated with and evaluated at $(X^*)_1^n$; C is a random variable depending on the data X_1^n, but being constant with respect to τ_c and hence irrelevant for model selection. For a justification of (4.4) see section 7. The expressions $-2\mathbb{E}_{\hat{P}_{\hat{c}_0}}[\log(\hat{P}_c^*((X^*)_1^n))]$ and $4\mathbb{E}_{\hat{P}_{\hat{c}_0}}[nV_n^*]$ are related to an information criterion playing the roles of a quantity similar to twice the negative log-likelihood $-2\log(\hat{P}_c(X_1^n))$ and of a penalty term, respectively. Note that by our definition of V_n there is a factor $1/2$ in its limiting distribution, see Theorem 3.3: the factor 4 in the penalty term here then corresponds to the more common factor 2 in the penalty term of AIC. Twice the negative log-likelihood plus the penalty, i.e., $-2\log(\hat{P}_c^*((X^*)_1^n) + 4\mathbb{E}_{\hat{P}_{\hat{c}_0}}[nV_n^*]$ is approximately unbiased (with respect to $\hat{P}_{\hat{c}_0}$) for $2\mathrm{EKLI}^*(\tau_c)$. In particular, the penalty term $4\mathbb{E}_{\hat{P}_{\hat{c}_0}}[nV_n^*]$ accounts for the fact that there is a plug-in bias in $\mathbb{E}_{\hat{P}_{\hat{c}_0}}[\log(\hat{P}_c^*((X^*)_1^n))]$, since the bootstrapped data $(X^*)_1^n$ is used in the estimate $\hat{P}_c^*$ and also as an argument in the log-likelihood $\log(\hat{P}_c^*(.))$: this is the same phenomenon as in $\log(\hat{P}_c(X_1^n)$ whose bias effect is corrected in the commonly known information criteria. In general, the term $4\mathbb{E}_{\hat{P}_{\hat{c}_0}}[nV_n^*]$ penalizing large models in $2\mathrm{EKLI}^*(\tau_c)$ is not converging to (the wrong constant) $2D(\tau_c)$ which would be the penalty term in AIC. The exception is when $\tau_c = \tau_{c_0}$ being the true model, see Remark 3.1. In our set-up, AIC is generally not a consistent criterion for minimizing $\mathrm{EKLI}(\tau_c)$.

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 models in $\mathcal{M}$ with context trees τ_c, the search over all these models can be computationally infeasible (even when considering only models in $\mathcal{M}$ having order $0 \le k \le K$ with K large). We focus here on the problem of finding the best model among the ones 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,

$$\hat{\tau}_0 \succ \hat{\tau}_1 \succ \ldots \succ \hat{\tau}_{\hat{m}-1} \succ \tau_{\hat{m}} = \tau_{root}, \tag{4.5}$$

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

$$K : X_1^n \to \hat{\tau}_K,\ \ \hat{\tau}_K \in \{\hat{\tau}_0, \ldots, \hat{\tau}_{\hat{m}-1}, \tau_{root}\}. \tag{4.6}$$

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 (4.6). We define the overall risk $R'(.)$ as

$$(4.7) R'(K) = \begin{cases} \mathbb{E}_{P_{c_0}}[L(Y_{n+1}, \varphi_K(\hat{c}_K(Y^n_{-\infty}), X_1^n))] \text{ for FPE with } L = L_2,\ \delta \\ \mathbb{E}_{P_{c_0}}[I_n(P_{c_0}, \hat{P}_{\hat{c}_K})] = \mathbb{E}_{P_{c_0}}[\int_{\mathcal{X}^n} \log\left(\frac{P_{c_0}(y_1^n)}{\hat{P}_{\hat{c}_K}(y_1^n)}\right) dP_{c_0}(y_1^n)] \text{ for EKLI} \end{cases},$$

where

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

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

$$K_{opt} = \mathrm{argmin}_K R'(K). \tag{4.8}$$

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 (4.6). 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

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

where $(Y^*)_1^{n+1}$, $(X^*)_1^n$ are as in (4.1) but with $\hat{P}_{\hat{c}_{K_0}}$ replacing the notation $\hat{P}_{\hat{c}_0}$.

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

$$\hat{K} = \mathrm{argmin}_K \mathrm{FPE}^*(K) \text{ or } \hat{K} = \mathrm{argmin}_K \mathrm{EKLI}^*(K). \tag{4.9}$$

Rigorous mathematical results for $\mathrm{FPE}^*(K)$, $\mathrm{EKLI}^*(K)$ or $\hat{K}$ in (4.9) 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.

We relate now the optimal cut-off K_{opt} or $\hat{K}$ in (4.8) and (4.9), respectively, to optimally pruned subtrees. Assume that we know the risk function $R(\tau_c)$ for all fixed models in $\mathcal{M}$ with structures τ_c. Optimality within the sequence of nested trees $\{\hat{\tau}_k\}_k$ in (4.5) then motivates the definition

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

The tree $\tilde{\tau}_{opt}$, which *depends on the data*, is called the 'optimally pruned sub-tree' with respect to the risk function $R(.)$, cf. Breiman et al. (1984, Chs. 3.3-3.4, 10). Note that it is only 'sub-optimal' in the sense that the optimization is over the computationally feasible class of nested trees in (4.5) rather than over all possible sub-trees of $\hat{\tau}_0 = \tau_{max}^t$. When the risk $R(.)$ is unknown, we can replace it by some estimate, in our case by the bootstrap estimate $R^*(.)$ (which can be evaluated at context trees) as in section 4 . However, the tree $\tilde{\tau}_{opt}$ might not be optimal with respect to some overall risk $R'(.)$ as in (4.7), treating $\hat{\tau}_k$ as random. Our algorithmic implementation as described above in the current section 4.1 (with bootstrap overall risk estimates $R^*(.)$ evaluated at cut-off values) can then be interpreted as aimed for 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 $\text{FPE}^*(K)$ 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 (4.6). 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 $\text{FPE}^*(K)$.

Instead of $\text{EKLI}(K)$ 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,

$$\text{NELL}(K) = -\int_{\mathcal{X}^n} \log(\hat{P}_{\hat{c}_K}(y_1^n)) dP_{c_0}(y_1^n),$$
$$\text{ENELL}(K) = \mathbb{E}_{P_{c_0}}[\text{NELL}(K)]. \tag{5.1}$$

The approximate calculation of $\text{ENELL}^*(K)$, analogous as for $\text{EKLI}^*(K)$ in Step 2 of the algorithm in section 4.1, can be done without integrating over $\mathcal{X}^n$. 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 (4.2), based on $\mathbf{X}_i^*$.

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

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 with respect to $y_1^n \sim \hat{P}_{\hat{c}_{K_0}}$ in NELL$^*(K)$, compare with formula (5.1), as well as of the expectation $\mathbb{E}_{\hat{P}_{\hat{c}_{K_0}}}[\text{NELL}^*(K)]$.

5.2 Simulations

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

(M1) Binary VLMC of order 8 ($\mathcal{X} = \{0,1\}$), as specified in Figure 5.1.

Figure 5.1 about here

(M2) 4-ary VLMC of order 2 ($\mathcal{X} = \{0,1,2,3\}$), as specified in Figure 5.2.

Figure 5.2 about here

We consider estimation of the different overall risks $R'(K)$ (FPE$_{L_2}(K)$, FPE$_\delta(K)$ and ENELL(K) as in (5.1)) with different initial cut-off values K_0 and the risks $R'(\hat{K})$ when plugging in the estimated cut-off parameter $\hat{K}$ in (4.9). 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 $\chi^2_{|\mathcal{X}|-1;0.9}/2$ and $\chi^2_{|\mathcal{X}|-1;0.8}/2$, 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.

Figure 5.3 about here

Figure 5.4 about here

Figures 5.3 and 5.4 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 (4.9) 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} [1]. All the sample versions are based on 100 simulations of the true process P_{c_0}.

Table 5.1 about here

Table 5.2 about here

Results are given in Tables 5.1 and 5.2 and are graphically displayed in Figures 5.3 and 5.4. The risk function ENELL is always standardized by the factor n^{-1}. We can summarize as follows.

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 $n = 1000$.

[1] 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 AM$_{P_{c_0}}(c_0(Y_{-\infty}^n))$, respectively. The oracle ENELL is $-\mathbb{E}_{P_{c_0}}[\log(P_{c_0}(Y_1^n))]$.

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 5.3 and 5.4 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., $|R'(\mathrm{argmin}_K \mathbb{E}[\hat{R}'(K)]) - R'(\mathrm{argmin}_K R'(K))|$ 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 models 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 terms, 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 obtain a method for choosing 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 related to optimal tree pruning in Breiman et al. (1984, Ch. 11.7) for CART with independent observations, but our approach takes the randomness of a pruned tree into account. As in risk estimation mentioned above, our method allows again a tuning 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.

The following questions about the alternative, competing predictive context algorithm for selection and estimation of VLMC's, briefly mentioned in section 1, remain open. What do asymptotic results tell for the predictive schemes? In particular, what kind of (sub-)optimality, in terms of an overall risk function R as in section 4.1, is achieved by the global-type predictive context algorithm in Bunton (1997)? And how does the latter compare with our (sub-)optimal solution for tuning the non-predictive context algorithm?

7 Proofs

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

We first remark that assumption (A) implies a Doeblin-type condition,

$$\sup_{A\subseteq\tau_{c_0};w,w'\in\tau_{c_0}} |p_Z^{(r)}(A,w)-p_Z^{(r)}(A,w')| < 1-\kappa \text{ for some } \kappa>0 \text{ and some } r\in\mathbb{N}, \tag{7.1}$$

where $p_Z^{(r)}(A,w)=\mathbb{P}[Z_r\in A|Z_0=w]$ denotes the r-step transition kernel of the state process $Z_t=c_0(X_{-\infty}^t)$, with $(X_t)_{t\in\mathbb{Z}}\sim P_{c_0}$. In particular, (7.1) implies a bound on the decay of the ϕ-mixing coefficients for P_{c_0},

$$\phi(i)\le(1-\kappa)^i,\ \ i\in\mathbb{N}. \tag{7.2}$$

Proof of Theorem 3.1. The decomposition $\text{FPE}_{L_2}(\tau_c)=S+B+V_n$ 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,

$$\begin{aligned}
V_n &= \mathbb{E}[\mathbb{E}[(\hat{\xi}-\xi)^2|c(Y_{-\infty}^n)]]\\
&= \mathbb{E}[Var(\hat{\xi}|c(Y_{-\infty}^n))]+\mathbb{E}[(\mathbb{E}[\hat{\xi}|c(Y_{-\infty}^n)]-\xi)^2|c(Y_{-\infty}^n)]=I_n+II_n. \qquad (7.3)
\end{aligned}$$

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

$$\begin{aligned}
\hat{P}_c(x|w) &= \frac{N(xw)}{N(w)}=\frac{n'^{-1}N(xw)}{P_{c_0}(w)}\\
&- \frac{n'^{-1}N(xw)}{P_{c_0}^2(w)}(n'^{-1}N(w)-P_{c_0}(w))+\frac{n'^{-1}N(xw)}{\tilde{P}^3(w)}(n'^{-1}N(w)-P_{c_0}(w))^2, \qquad (7.4)
\end{aligned}$$

where $|\tilde{P}(w)-P_{c_0}(w)|\le|n'^{-1}N(w)-P_{c_0}(w)|$, and $N(.)$ as in (2.2).
By assumption (A), which ensures the geometric ϕ-mixing property, see (7.2), we get

$$\begin{aligned}
&n^{1/2}(n'^{-1}N(w)-P_{c_0}(w))\Rightarrow\mathcal{N}(0,\sigma^2(w)),\\
&\sigma^2(w)=\sum_{k=-\infty}^{\infty}Cov(1_{[X_0^{m-1}=w]},1_{[X_k^{k+m-1}=w]}),\ \ m=|w|, \qquad (7.5)
\end{aligned}$$

and

$$nCov(n'^{-1}N(xw), n'^{-1}N(w)) \to \tau^2(xw),$$

$$\tau^2(xw) = \sum_{k=-\infty}^{\infty} Cov(1_{[X_0^m = xw]}, 1_{[X_k^{k+m-1}=w]}),\ m = |w|. \tag{7.6}$$

Using (7.5), (7.6) 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 $\tilde{P}(w) > 0\ a.s.(P_{c_0})$, $0 \le n'^{-1}N(xw) \le 1$ and by the geometric ϕ-mixing property of P_{c_0} given in (7.2), together with the boundedness of indicator functions) we get

$$n\mathbb{E}[\hat{P}_c(x|w) - P_{c_0}(x|w)|w] = -\frac{1}{P_{c_0}^2(w)}\tau^2(xw) + \frac{P_{c_0}(x|w)}{P_{c_0}^2(w)}\sigma^2(w) + o(1). \tag{7.7}$$

With (7.7) and the finiteness of τ_c we get

$$II_n = \mathbb{E}[(\mathbb{E}[\hat{\xi}|c(Y_{-\infty}^n)] - \xi)^2|c(Y_{-\infty}^n)] = O(n^{-2}). \tag{7.8}$$

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

$$nVar(\hat{\xi}|w) = \sum_{x_1,x_2 \in \mathcal{X}} x_1 x_2 nCov(\frac{N(x_1w)}{N(w)}, \frac{N(x_2w)}{N(w)}),$$

and using an expansion similar as in (7.4) we obtain with $n' = n - |w|$,

$$nVar(\hat{\xi}|w) = \sum_{x_1,x_2 \in \mathcal{X}} x_1 x_2 \frac{1}{P_{c_0}^2(w)} nCov(n'^{-1}N(x_1w), n'^{-1}N(x_2w)) + o(1).$$

Similar to (7.6) we then get with $m = |w|$,

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

Thus, by integrating over $w = c(Y_{-\infty}^n)$, $nI_n = C(\tau_c, P_{c_0}) + o(1)$. This, together with (7.3) and (7.8) 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

$$\begin{aligned} |V_n| &= |\mathbb{E}[\mathbb{E}[1_{[Y_{n+1} \ne \mathrm{AM}_{P_{c_0}}(c(Y_{-\infty}^n))]} - 1_{[Y_{n+1} \ne \varphi(c(Y_{-\infty}^n), X_1^n)]}|c(Y_{-\infty}^n)]]| \\ &\le \mathbb{E}[\mathbb{E}[1_{[\varphi(c(Y_{-\infty}^n), X_1^n) \ne \mathrm{AM}_{P_{c_0}}(c(Y_{-\infty}^n))]}|c(Y_{-\infty}^n)]]. \end{aligned} \tag{7.9}$$

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

$$\begin{aligned}
\mathbb{P}[\varphi(w, X_1^n) \neq \mathrm{AM}_{P_{c_0}}(w)|w] &\leq \mathbb{P}[\cup_{x\in\mathcal{X}}\{|\hat{P}_c(x|w) - P_{c_0}(x|w)| > \varepsilon/2\}|w] \\
&\leq \sum_{x\in\mathcal{X}} \mathbb{P}[|\hat{P}_c(x|w) - P_{c_0}(x|w)| > \varepsilon/2|w]. \qquad (7.10)
\end{aligned}$$

Similarly as in (7.4) we get with $n' = n - |w|$,

$$\begin{aligned}
&\hat{P}_c(x|w) - P_{c_0}(x|w) \\
=\ & \frac{1}{P_{c_0}(w)}(n'^{-1}N(xw) - P_{c_0}(xw)) - \frac{n'^{-1}N(xw)}{\tilde{P}^2(w)}(n'^{-1}N(w) - P_{c_0}(w)) \\
=\ & I_n - II_n, \qquad (7.11)
\end{aligned}$$

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,

$$|I_n| \leq \varepsilon/6 P_{c_0}(x|w) \leq \varepsilon/6 \text{ on } D_n^C(x,w). \qquad (7.12)$$

For the second term II_n in (7.11), 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,

$$|II_n| \leq \varepsilon/3 \text{ on } D_n^C(x,w) \cap E_n^C(w). \qquad (7.13)$$

Therefore, by (7.11)-(7.13),

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

Thus, by formula (7.9) and (7.10),

$$\begin{aligned}|V_n| &\le \sum_{w\in\tau_c}\sum_{x\in\mathcal{X}}(\mathbb{P}[D_n(x,w)]+\mathbb{P}[E_n(w)])P_{c_0}(w)\\ &\le |\mathcal{X}|(\max_{x\in\mathcal{X},w\in\tau_c}\mathbb{P}[D_n(x,w)]+\max_{w\in\tau_c}\mathbb{P}[E_n(w)]).\end{aligned} \tag{7.14}$$

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

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

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), using the bound on the ϕ-mixing coefficients in (7.2). Thus, in the notation of Doukhan's Proposition 2, $k_n \le C(\kappa)\log(n')$, $C(\kappa) > 0$ 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. Also, for both type of sets we use $\sigma = A(\log(n'))^{1/2}$ ($A > 0$ a constant) in the notation of Doukhan. Then, for A sufficiently large, the restriction in Doukhan $0 \le x \le \frac{\sigma\sqrt{n'}}{8bk_n}$ holds. Note that $n' \ge n-k_c$ and by assumption (B), $P_{c_0}(xw) \ge \pi$ for all xw. Then, Proposition 2 in Doukhan (1994, Ch. 1.4.2) yields for $n' > 30/\varepsilon$, i.e., for n sufficiently large,

$$\max_{x\in\mathcal{X},w\in\tau_c}\mathbb{P}[D_n(x,w)] \le \tilde{C}_1\exp(-C_2(\kappa)\varepsilon^2\pi^2(n')/\log(n')),$$

where $\tilde{C}_1, C_2 = C_2(\kappa) > 0$, and an even better bound (a bigger constant $C_2 = C_2(\kappa)$) applies for $\max_{w\in\tau_c}\mathbb{P}[E_n(w)]$.
By setting $C_1 = 2\tilde{C}_1$ these bounds, together with (7.14) complete the proof. □

Proof of Theorem 3.3. We decompose

$$\mathrm{KLI}(\tau_c)/n = B_n + V_n,\ nV_n = \int_{\mathcal{X}^n}\log\left(\frac{\bar{P}_c(y_1^n)}{\hat{P}_c(y_1^n)}\right)dP_{c_0}(y_1^n). \tag{7.15}$$

It is then helpful to parameterize the probability measures on $\mathcal{X}^\infty$ 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 $\mathcal{X} = \{0,\ldots,|\mathcal{X}|-1\}$: then, these transition probabilities are indexed as

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

$$(\theta_0)_{wx} = P_{c_0}(x|w) = P_{c_0}(xw)/P_{c_0}(w),\ w \in \tau_{c_0}.$$

As in standard maximum likelihood theory we expand

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

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 $\mathbb{E}[U_{(c,\bar\theta)}(Y_1^n)] = \int_{\mathcal{X}^n} U_{(c,\bar\theta)}(y_1^n)dP_{(c_0,\theta_0)}(y_1^n) = 0$ we have by (7.15),

$$nV_n = -1/2(\hat\theta-\bar\theta)^T\int_{\mathcal{X}^n} H_{(c,\tilde\theta)}(y_1^n)dP_{(c_0,\theta_0)}(\hat\theta-\bar\theta). \tag{7.16}$$

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

$$U_{(c,\theta)}(X_1^n) = \sum_{t=k_c+1}^n \tilde U_{(c,\theta)}(X_{t-k_c}^t) + o_P(1),$$
$$\tilde U_{(c,\theta)}(X_{t-k_c}^t) = \frac{\partial}{\partial\theta}\log(P_{(c,\theta)}(X_t|c(X_{t-k_c}^{t-1})) = \frac{\partial}{\partial\theta}\log(\theta)_{c(X_{t-k_c}^{t-1}),X_t},$$

where k_c is the order of $c(.)$ (the depth of τ_c).
At $\bar\theta$ and for the component index wx,

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

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

$$n^{-1/2}\sum_{t=k_c+1}^n \tilde U_{(c,\bar\theta)}(X_{t-k_c}^t) \Rightarrow \mathcal{N}(0,F(c,\bar\theta)),$$
$$F(c,\bar\theta) = \sum_{m=-\infty}^{\infty}\mathbb{E}[\tilde U_{(c,\bar\theta)}(X_{-k_c}^0)\tilde U_{(c,\bar\theta)}^T(X_{m-k_c}^m)]. \tag{7.17}$$

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 $F(c,\bar\theta) = \mathbb{E}[\tilde U_{(c,\bar\theta)}(X_{-k_c}^0)\tilde U_{(c,\bar\theta)}(X_{-k_c}^0)^T]$.
The Hessian matrix in (7.16) is of the form

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

Thus, the limit of the expected value is given by

$$
\begin{aligned}
J(c,\bar{\theta}) &= \lim_{n\to\infty} n^{-1} \int_{\mathcal{X}^n} H_{(c,\bar{\theta})}(y_1^n) dP_{(c_0,\theta_0)}(y_1^n) \\
&= -\delta_{w_1 w_2}(\delta_{x_1 x_2} \frac{1}{\bar{\theta}_{w_1 x_1}} + \frac{1}{1-\sum_{r=0}^{|\mathcal{X}|-1} \bar{\theta}_{w_1 r}}) P_{(c_0,\theta_0)}(w_1).
\end{aligned} \tag{7.18}
$$

It is straightforward to show $\tilde{\theta} = \bar{\theta} + o_P(1)$. We then get for the expression in (7.16),

$$
\int_{\mathcal{X}^n} n^{-1} H_{(c,\tilde{\theta})}(y_1^n) dP_{(c_0,\theta_0)}(y_1^n) = J(c,\bar{\theta}) + o_P(1). \tag{7.19}
$$

Also, by standard arguments for MLE, using (7.17), (7.18) and the mixing property of P_{c_0}, we get

$$
n^{1/2}(\hat{\theta} - \bar{\theta}) \Rightarrow -J(c,\bar{\theta})^{-1} F(c,\bar{\theta})^{1/2} Z,\ Z \sim \mathcal{N}_{D(\tau_c)}(0, I). \tag{7.20}
$$

Thus, by (7.16), (7.19) and (7.20) we get

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

Since $\bar{\theta}$ is a function of P_{c_0} on τ_c, and since the quantities $F(.,.)$ in (7.17) and $J(.,.)$ in (7.18) 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 (7.15) 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 $nV_n \Rightarrow 1/2\chi^2_{D(\tau_{c_0})}$.

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 (A). Assume that the cut-off $K_n > (2|\mathcal{X}| + 4)\log(n)$ 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* $x_1^m \in \mathcal{X}^m$ $(m \in \mathbb{N})$,

(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 (A) and also (7.1) with $\kappa/2$ replacing the value κ assumed for P_{c_0}.

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 $\phi_{\hat{P}_{\hat{c}_0}}(k) \le (1-\kappa/2)^k$ 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 (A), implying (7.1) and (7.2) 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 (B) with P_{c_0} replaced by $\hat{P}_{\hat{c}_0}$ holds in probability (with $\varepsilon/2$ replacing the value ε assumed for P_{c_0}). 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 $\mathrm{KLI}^*(\tau_c)/n = B^*_n + V^*_n$ 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 nV^*_n. By Lemma 7.1(iii) we can carry out the same steps as in the proof of Theorem 3.3 to obtain

$$\mathbb{P}_{\hat{P}_{\hat{c}_0}}[nV^*_n \le x] = \mathbb{P}[1/2 Z^T \Sigma(\tau_c, \hat{P}_{\hat{c}_0}) Z \le x | \hat{P}_{\hat{c}_0}] + o_P(1), \ x \in \mathbb{R},$$
$$\Sigma(\tau_c, \hat{P}_{\hat{c}_0}) = F(c, \bar{\theta}^*)^{1/2} J(c, \bar{\theta}^*)^{-1} F(c, \bar{\theta}^*)^{1/2}, \tag{7.21}$$

with $F(.,.)$ as in (7.17) and $J(.,.)$ as in (7.18), but with $P_{(\hat{c}_0, \hat{\theta}_0)}$ instead of $P_{(c_0,\theta_0)}$: here $(\bar{\theta}^*)_{wx} = \hat{P}_{\hat{c}_0}(xw)/\hat{P}_{\hat{c}_0}(w)$, $w \in \tau_c$.
By Lemma 7.1(i)-(ii) we then get

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

and thus $\Sigma(\tau_c, \hat{P}_{\hat{c}_0}) = \Sigma(\tau_c, P_{c_0}) + o_P(1)$. Together with (7.21), this completes the proof. □

Proof of formula (4.4). Write

$$\mathrm{KLI}^*(\tau_c) = C/2 - \int_{\mathbb{R}^n} \log(P_{(c,\hat{\theta}^*)}(y_1^n)) dP_{(\hat{c}_0,\hat{\theta})},$$
$$C = 2 \int_{\mathbb{R}^n} \log(P_{(\hat{c}_0,\hat{\theta})}(y_1^n)) dP_{(\hat{c}_0,\hat{\theta})}(y_1^n). \tag{7.22}$$

By definition of V_n^*,

$$(7.23) - \int_{\mathbb{R}^n} \log(P_{(c,\hat{\theta}^*)}(y_1^n)) dP_{(\hat{c}_0,\hat{\theta})}(y_1^n) = - \int_{\mathbb{R}^n} \log(P_{(c,\bar{\hat{\theta}})}(y_1^n)) dP_{(\hat{c}_0,\hat{\theta})}(y_1^n) + nV_n^*,$$

where $P_{(c,\bar{\hat{\theta}})}$ is given by the transition probabilities $\bar{\hat{\theta}}_{wx} = P_{(\hat{c}_0,\hat{\theta})}(xw)/P_{(\hat{c}_0,\hat{\theta})}(w)$ for $w \in \tau_c$.
On the other hand, expanding $\log(P_{(c,\bar{\hat{\theta}})}((X^*)_1^n))$ around $\hat{\theta}^*$, using that $\frac{d}{d\theta}\log(P_{(c,\theta)}((X^*)_1^n))|_{\theta=\hat{\theta}^*} = 0$ and taking expectations with respect to $P_{(\hat{c}_0,\hat{\theta})}$, we obtain

$$\begin{aligned} & \int_{\mathbb{R}^n} \log(P_{(c,\bar{\hat{\theta}})}((X^*)_1^n)) dP_{(\hat{c}_0,\hat{\theta})}((X^*)_1^n) \\ \approx\ & \mathbb{E}_{\hat{P}_{\hat{c}_0}}[\log(P_{(c,\hat{\theta}^*)}((X^*)_1^n))] - \mathbb{E}_{\hat{P}_{\hat{c}_0}}[nV_n^*], \end{aligned} \tag{7.24}$$

where the approximate sign '$\approx$' is justified by convergence in distribution of $-nV_n^*$ and the second-order remainder in the expansion of $\log(P_{(c,\bar{\hat{\theta}})}((X^*)_1^n))$ to the same limit (compare also with formula (7.16)) and assuming a uniform integrability argument that the difference of the corresponding expected values goes to zero.
Thus, by (7.22)-(7.24) we get

$$2\mathrm{EKLI}^*(\tau_c) \approx C - 2\mathbb{E}_{\hat{P}_{\hat{c}_0}}[\log(\hat{P}_c^*((X^*)_1^n))] + 4\mathbb{E}_{\hat{P}_{\hat{c}_0}}[nV_n^*],$$

which completes the proof of (4.4). □

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. Moreover, constructive comments by both referees helped to improve the presentation of the results.

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] Akaike, H. (1973). Information theory and the maximum likelihood principle. In 2nd International Symposium on Information Theory (Eds. B.N. Petrov and F. Csàki), pp. 267-281. Akademiai Kiàdo, Budapest.

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

[5] Bühlmann, P. (1998). Efficient and adaptive post-model-selection estimators. To appear in J. Statist. Plann. Infer.

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

[7] Bunton, S. (1997). A percolating state selector for suffix-tree context models. In Proceedings of the 1997 Data Compression Conference, Snowbird, Utah (Eds. J.A. Storer and M. Cohn), pp. 32-41. IEEE Computer Society Press.

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

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

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

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

[12] Merhav, N., Gutman, M. and Ziv, J. (1989). On the estimation of the order of a Markov chain and universal data compression. IEEE Trans. Inform. Theory **IT-35** 1014-1019.

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

[14] Rissanen, J. (1986). Complexity of strings in the class of Markov sources. IEEE Trans. Inform. Theory **IT-32** 526-532.

[15] Rissanen, 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.

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

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

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

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

[20] Weinberger, M.J. and Feder, M. (1994). Predictive stochastic complexity and model estimation for finite-state processes. J. Statist. Plann. Infer. **39** 353-372.

[21] Weinberger, M.J., Lempel, A. and Ziv, J. (1992). A sequential algorithm for the universal coding of finite memory sources. IEEE Trans. Inform. Theory **IT-38** 1002-1014.

[22] Weinberger, M.J., Rissanen, 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.

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

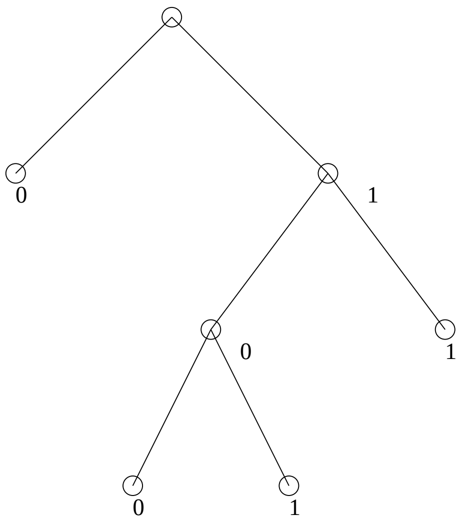

Figure 2.1: Context tree τ_c in Example 2.1.

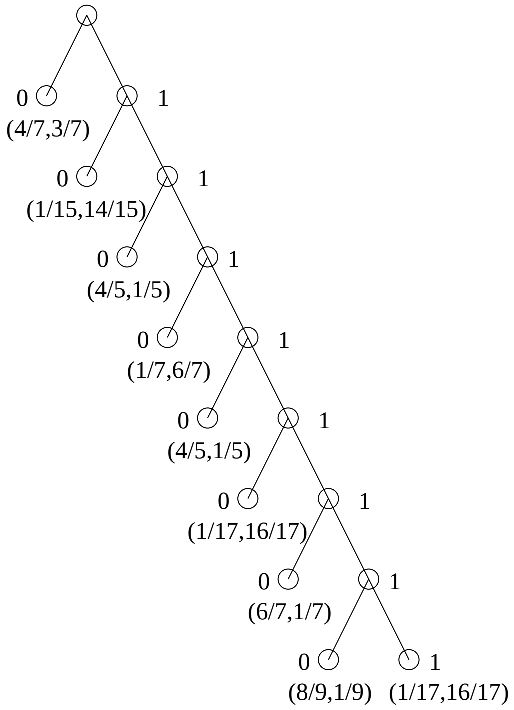

Figure 5.1: Context tree and transition probabilities for (M1).

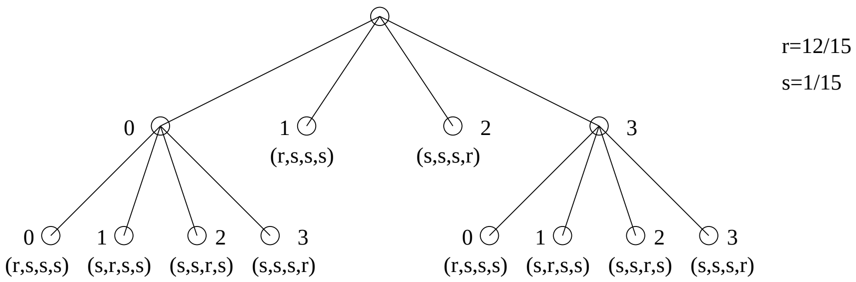

Figure 5.2: Context tree and transition probabilities for (M2).

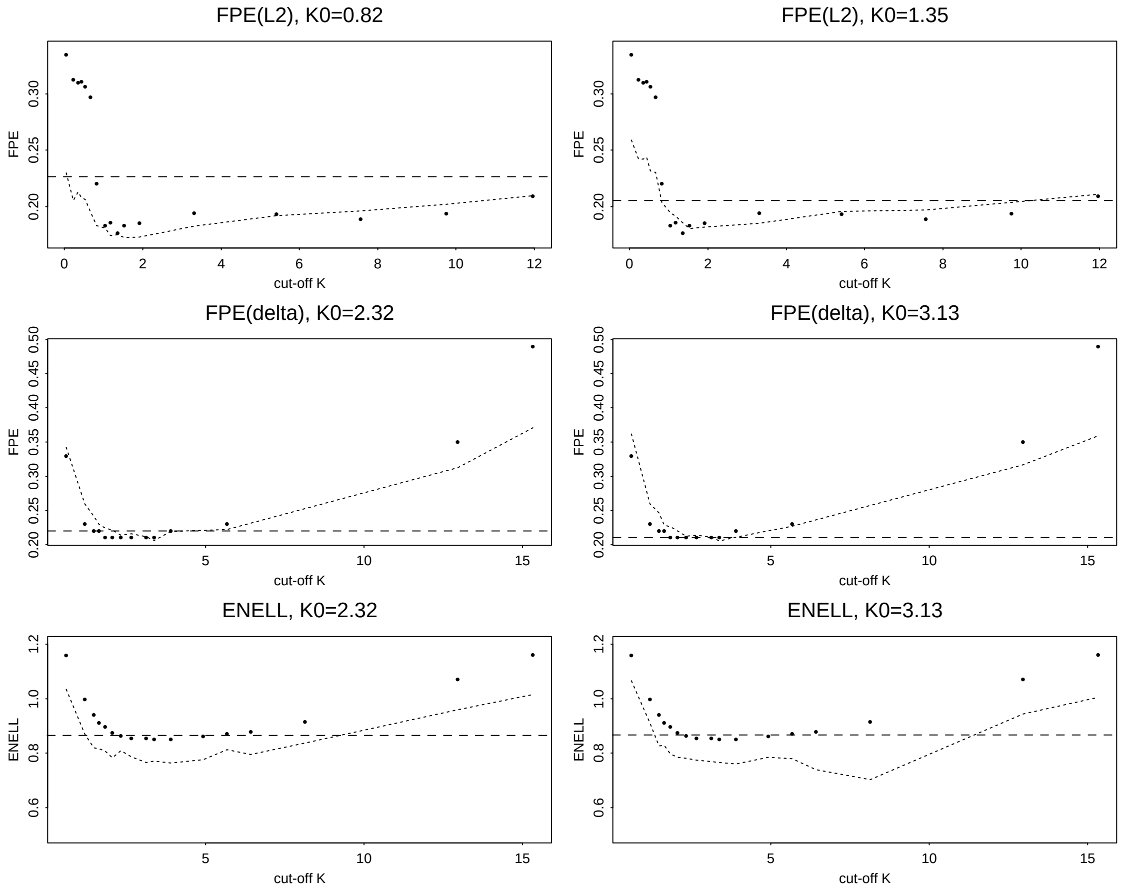

Figure 5.3: Risks for sample size $n = 200$. Model (M1) for FPE_{L_2}, model (M2) for FPE_δ and ENELL, respectively. Dots: $R'(K)$; dotted line: $\mathbb{E}[\hat{R}'(K)]$; dashed line: $\mathbb{E}[R'(\hat{K})]$.

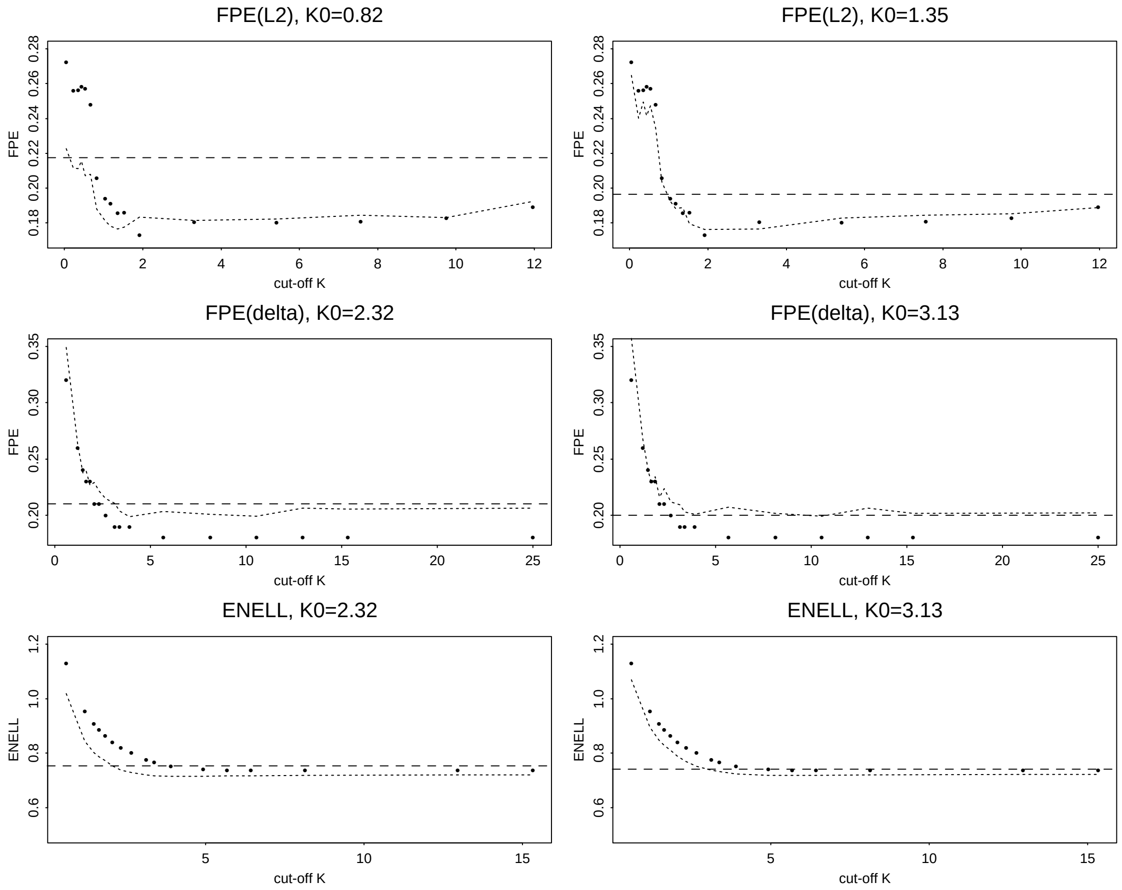

Figure 5.4: Risks for sample size $n = 1000$. Model (M1) for FPE_{L_2}, model (M2) for FPE_δ and ENELL, respectively. Dots: $R'(K)$; dotted line: $\mathbb{E}[\hat{R}'(K)]$; dashed line: $\mathbb{E}[R'(\hat{K})]$.

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 $n = 1000$.