Dynamic Adaptive Partitioning for Nonlinear Time Series

by

Peter Bühlmann

Research Report No. 84
April 1998

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Dynamic Adaptive Partitioning for Nonlinear Time Series

Peter Bühlmann
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

April 1998

Abstract

We propose a dynamic adaptive partitioning scheme for nonparametric analysis of stationary nonlinear time series with values in $\mathbb{R}^d$ $(d \geq 1)$. We use information from past values to construct adaptive partitioning in a *dynamic* fashion which is then different from the more common static schemes in the regression set-up.

The idea of dynamic partitioning is novel. We make it constructive by proposing an approach based on quantization of the data and adaptively modelling partition cells with a parsimonious Markov chain. The methodology is formulated in terms of a model class, the so-called quantized variable length Markov chains (QVLMC).

We discuss when and why such a QVLMC partitioning scheme is in some sort natural and canonical. We explore asymptotic properties and give some numerical results which reflect the finite sample behavior.

Key words and phrases. Conditional heteroscedasticity, Context algorithm, Markov chain, Multivariate time series, ϕ-mixing, Prediction, Quantization, Stationary sequence, Tree model.

Short title: Dynamic adaptive partitioning

1 Introduction

We propose a nonparametric analysis for nonlinear stationary time series which is based on adaptive partitioning. The nonparametric approach is appropriate if there is essentially no pre-knowledge about the underlying dynamic characteristics of the process and if the sample size of the observed data is not too small. Such situations arise nowadays in quite many applications. We refer to Tjøstheim (1994) for a review of nonparametric (and parametric) models and to Priestley (1988) and Tong (1990) for nonlinear parametric models.

Nonparametric methods which are able to adapt to local sparseness of the data are very attractive, they are often substantially better than non-adaptive procedures because of the curse of dimensionality. In nonparametric modelling of nonlinear phenomena, estimation of the mean as a function of predictor variables (the regression function) with adaptive partitioning schemes has attracted much attention, cf. Breiman et al. (1984) with CART, Friedman (1991) with MARS or Gersho & Gray (1992) for an overview with a more information theoretical point of view. Some of the partitioning schemes have been studied also for the case of stationary time series, cf. Lewis & Stevens (1991) with MARS or Nobel (1997) for general distortion properties of partitioning schemes.

But none of these adaptive partitioning schemes is using the simple fact that in case of a time series, the partition cells themselves have typically a dynamic characteristic. Consider a stationary real-valued p-th order Markov chain Y_t ($t \in \mathbb{Z}$) with state vector $S_{t-1} = (Y_{t-1}, \ldots, Y_{t-p})$ being the first p lagged variables. Adaptive partitioning typically yields models of the form $\mathbb{E}[Y_t|S_{t-1}] = \sum_{j=1}^{J} c_j 1_{[S_{t-1} \in R_j]}$ with $\{R_j; j = 1, \ldots, J\}$ a partition of the state space $\mathbb{R}^p$. (MARS (Friedman, 1991) uses splines instead of step functions). This is the common model in the regression set-up with independent errors; the various schemes differ by adaptively producing different partitions. But for the time series case we observe and make use of the following two facts.

(1) Y_t is the first component of the next state vector S_t.

(2) $V_{t-1} = \sum_{j=1}^{J} R_j 1_{[S_{t-1} \in R_j]}$ ($t \in \mathbb{Z}$) is a stochastic process with values in $\{R_j;\ j = 1, \ldots, J\}$. Note that $S_{t-1} \in V_{t-1}$ for all $t \in \mathbb{Z}$. Given $V_1, \ldots, V_{t-1}$ (or $Y_1, \ldots, Y_{t-1}$), we can learn about a future partition cell V_t.

These facts (1) and (2) simply say that we can learn partially about Y_t via V_t (being the partition cell of which S_t will be an element) from the partition cell process $V_1, \ldots, V_{t-1}$ or the data $Y_1, \ldots, Y_{t-1}$. The novel approach here is to additionally *model* the partition cell process $(V_t)_{t\in\mathbb{Z}}$, thus 'making dependence our friend' for adaptive partitioning. We incorporate this idea by quantization of the real-valued (or multivariate $\mathbb{R}^d$-valued) data which then yields an adaptive partition cell process as in the fact (2) above, being modelled with a parsimonious Markov chain. This explains also the expression 'dynamic adaptive partitioning' in the title. It is possible to combine the quantization operation and the Markov modelling of the partition cell process in a properly defined model class for stationary, ergodic time series with values in $\mathbb{R}^d$ ($d \geq 1$), the so-called quantized variable length Markov chains (QVLMC). We argue in sections 2.6, 2.7 and 3.1 why this model class and its adaptive estimation turns out to be some sort of canonical in the context of dynamic adaptive partitioning.

The paper is organized as follows. In section 2 we describe the above mentioned model class with its properties, in section 3 we discuss adaptive partitioning and estimation of the models, in section 4 we give some results for asymptotic inference, in section 5 we discuss the issue about model selection, in section 6 we present some numerical examples and we state some conclusions in section 7. All proofs are deferred to an Appendix.

2 The QVLMC model

Our general strategy to find and fit a nonlinear time series model is to quantize the data first and then use an adaptively estimated parsimonious Markov model for the quantized series. To make our strategy successful having a finite amount of data, we need a good technique for choosing the amount of quantization and a good model for the quantized series. Both issues are addressed in section 5.

In general, we assume that the data $Y_1, \ldots, Y_n$ is an $\mathbb{R}^d$-valued stationary time series. Denote by

$$q : \mathbb{R}^d \to \mathcal{X} = \{0, 1, \ldots, N-1\} \tag{2.1}$$

a quantizer of $\mathbb{R}^d$ into a categorical set $\mathcal{X} = \{0, 1, \ldots, N-1\}$. To be precise, q describes the discrete structure of a quantizer and does not assign a representative value in $\mathbb{R}^d$, i.e., a so-called word in the code book. The quantizer q gives rise to a partition of $\mathbb{R}^d$,

$$\begin{aligned} &\mathbb{R}^d = \cup_{x \in \mathcal{X}} I_x,\ I_x \cap I_y = \emptyset\ (x \neq y), \\ &y \in I_{q(y)} \text{ for all } y \in \mathbb{R}^d. \end{aligned} \tag{2.2}$$

2.1 VLMC for categorical variables

Consider a stationary process $(X_t)_{t \in \mathbb{Z}}$ with values in a finite categorical space $\mathcal{X} = \{0, 1, \ldots, N-1\}$ as in (2.1). We will see in section 2.2 that X_t will play the role of a quantized variable $Y_t \in \mathbb{R}^d$, i.e., $X_t = q(Y_t)$ with q as in (2.1).

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. First, we define the variable length Markov chains. Such kind of models have been introduced in information theory as tree models, FSMX models or finite-memory sources. More motivation is given in Rissanen (1983), Weinberger et al. (1995) or Bühlmann & Wyner (1997).

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

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

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

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

Then $c(.)$ is called a context function of order p, and $(X_t)_{t\in\mathbb{Z}}$ is called a stationary variable length Markov chain (VLMC) of order p.

We sometimes identify a VLMC $(X_t)_{t\in\mathbb{Z}}$ with its probability distribution P_c on $\mathcal{X}^{\mathbb{Z}}$. Also, we often write $P_c(x_i^j) = \mathbb{P}[X_i^j = x_i^j]$ and $P_c(x_j|x_i^{j-1}) = \mathbb{P}[X_j = x_j|X_i^{j-1} = x_i^{j-1}]$ $(i < j)$ for $(X_t)_{t\in\mathbb{Z}} \sim P_c$.

Clearly, a VLMC of order p is a Markov chain of order p, now having a *memory of variable length* ℓ. By requiring stationarity, a VLMC is thus completely specified by its transition probabilities $P_c(x_1|c(x_{-\infty}^0))$, $x_{-\infty}^1 \in \mathcal{X}^\infty$. Many context functions $c(.)$ 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.

A VLMC is a tree structured model with a root node on top, from which the branches are growing downwards, so that every internal node has at most $N = |\mathcal{X}|$ offsprings. Then, each value of a context function $c(.)$ can be represented as a branch (or terminal node) of such a tree. The context $w = c(x_{-\infty}^0)$ 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 $x_{-\ell(x_{-\infty}^0)+1}$.

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

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

can be represented by the tree $\tau = \tau_c$, see Figure 2.1.
A 'growing to the left' sub-branch represents the symbol 0 *and vice versa for the symbol* 1.

Note that such context trees do not have to be complete, i.e., every internal node does not need to have exactly $N = |\mathcal{X}|$ offsprings.

Definition 2.3 *Let $c(.)$ be a context function of a stationary VLMC of order p. The context tree τ and terminal node context tree τ^T are defined as*

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

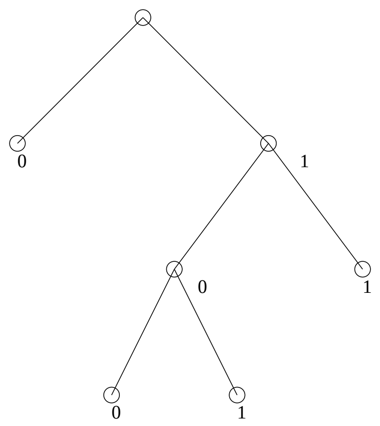

Figure 2.1: Context tree τ_c from example 2.1.

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 a VLMC with context function $c(.)$. An internal node with $b < N = |\mathcal{X}|$ offsprings can be implicitly thought to be complete by adding one complementary offspring, lumping the $N - b$ non-present nodes together to a single new terminal node w_{new} representing a single state in τ_c.

2.2 QVLMC for $\mathbb{R}^d$-valued variables

Let $q,\ \mathcal{X},\ I_x$ be as in (2.1) and (2.2), respectively. Assume that

$$(X_t)_{t\in\mathbb{Z}} \text{ is an } \mathcal{X}\text{-valued finite order variable length Markov chain (VLMC)}, \tag{2.3}$$

as described in Definition 2.2. Given $X_t = x$, define Y_t independently of Y_s $(s \neq t)$ and X_s $(s \neq t)$,

$$\begin{aligned} &Y_t \sim f_x(y)dy \text{ given } X_t = x, \\ &supp(f_x) \subseteq I_x, \text{ for all } x \in \mathcal{X}, \end{aligned} \tag{2.4}$$

where $f_x(.)$ is a d-dimensional density with respect to Lebesgue.

Definition 2.4 *The process $(Y_t)_{t\in\mathbb{Z}}$ defined by (2.3) and (2.4) is called a stationary quantized variable length Markov chain, denoted by QVLMC.*

A QVLMC has the property that its quantized values (with the correct quantizer q) form a VLMC, i.e., $(q(Y_t))_{t\in\mathbb{Z}} = (X_t)_{t\in\mathbb{Z}}$ is a VLMC. It is sometimes useful to think of the underlying VLMC $(X_t)_{t\in\mathbb{Z}}$ also as a VLMC $(I_{q(Y_t)})_{t\in\mathbb{Z}}$ which has as values the sets $I_{q(Y_t)}$ being most often intervals.

A QVLMC $(Y_t)_{t\in\mathbb{Z}}$ is a stationary $\mathbb{R}^d$-valued Markov chain with a memory of variable length because

$$\mathbb{P}[Y_t \le y|Y_{-\infty}^{t-1}] = \sum_{x\in\mathcal{X}} \int_{-\infty}^{y} f_x(z)dz\mathbb{P}[X_t = x|c(X_{-\infty}^{t-1})],\ y \in \mathbb{R}^d,\ X_s = q(Y_s)$$

('$\le$' is defined componentwise), and thus,

$$\mathbb{P}[Y_t \le y|Y_{-\infty}^{t-1}] = \mathbb{P}[Y_t \le y|c(X_{-\infty}^{t-1})],\ y \in \mathbb{R}^d,\ X_s = q(Y_s).$$

The minimal state space of $(Y_t)_{t\in\mathbb{Z}}$ is the same as for $(X_t)_{t\in\mathbb{Z}}$, namely τ_c (see Definition 2.3). The dynamical character of a QVLMC is given by the probabilistic model for the quantized series, namely the underlying VLMC.

The class of QVLMC's is broader than the nonparametric autoregressive models

$$Y_t = m(Y_{t-p}^{t-1}) + Z_t\ (t \in \mathbb{Z},\ p \in \mathbb{N}) \tag{2.5}$$

with i.i.d. innovation noise $(Z_t)_{t\in\mathbb{Z}}$, $Y_t \in \mathbb{R}^d$ and $m(.)$ the nonparametric autoregressive function. In contrast to this model class in (2.5), the QVLMC's are flexible enough to approximate any stationary $\mathbb{R}^d$-valued process, see Theorem 2.1 in section 2.4. We view the QVLMC's as an important addition of a different class of models for stationary nonlinear time series.

For the univariate QVLMC model, the quantizer $q : \mathbb{R} \to \mathcal{X} = \{0, 1, \ldots, N-1\}$ in (2.1) is usually chosen with an interval geometry, see formula (3.1). That is, the sets I_x $(x \in \mathcal{X})$ in (2.2) are disjoint intervals in $\mathbb{R}$.

For the multivariate QVLMC model, the quantizer is $q : \mathbb{R}^d \to \mathcal{X} = \{0, 1, \ldots, N-1\}$. General vector quantization is less interpretable than scalar quantization, particularly in terms of individual series. We propose, but do not require, scalar quantization of different individual time series,

$$q_j : \mathbb{R} \to \mathcal{X}_j = \{0, 1, \ldots, N_j\},\ j = 1, \ldots, d, \tag{2.6}$$

which can be combined to a quantizer,

$$\begin{aligned} &q : \mathbb{R}^d \to \mathcal{X},\ q(Y_t) = (q_1(Y_{1,t}), \ldots, q_d(Y_{d,t})),\ Y_t = (Y_{1,t}, \ldots, Y_{d,t}),\\ &\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d, \end{aligned} \tag{2.7}$$

and $\mathcal{X}$ is labelled (arbitrarily) by $0, 1, \ldots, N-1$ with $N = N_1 \cdots N_d$: $\mathcal{X}$ is the product space of the different quantized values from the individual time series.

The class of multivariate QVLMC models is very flexible. Already the individual quantizers in (2.6) allow different degrees of 'resolution', i.e., different numbers N_j for q_j. Choosing a general vector quantization $q : \mathbb{R}^d \to \mathcal{X}$ can be done in very many different ways: this can be an advantage in a certain application, but generally it is easier to stick with individual quantization as in (2.6). The flexibility of QVLMC's also allows to model multidimensional time series data with some real-valued and some categorical components. Nonlinear multivariate time series modelling has not received very much attention so far. One main difficulty is to deal with the complexity of the data in an appropriate way. The recursive partitioning idea with TSMARS has been explored to the so-called semi-multivariate case (Lewis and Stevens, 1991) where one time series is the response series and the others are covariate series. Our approach has the potential to be useful for the pure multivariate case.

2.3 Ergodicity of QVLMC's

The dynamic property of a QVLMC is given by the VLMC model of the quantized series $X_t = q(Y_t)$. Since in the QVLMC model the variables Y_t given $(X_s)_{s \in \mathbb{Z}}$ are independent and depend only on their quantized values X_t, stationarity and ergodicity of $(Y_t)_{t \in \mathbb{Z}}$ is inherited from the VLMC $(X_t)_{t \in \mathbb{Z}}$. (Note that this statement is meant to be unconditional on $(X_t)_{t \in \mathbb{Z}}$. This is appropriate here, because the values X_t are non-hidden). A sufficient condition for ergodicity is then implied by a Doeblin-type condition, stationarity is already implicitly assumed by our Definitions 2.2 and 2.4.

(A) The underlying quantized finite order VLMC $(X_t)_{t \in \mathbb{Z}} \sim P_c$ on $\mathcal{X}^{\mathbb{Z}}$ satisfies,

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

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

Proposition 2.1 *Let $(Y_t)_{t \in \mathbb{Z}}$ be a QVLMC as given in Definition 2.4, satisfying condition (A). Then, $(Y_t)_{t \in \mathbb{Z}}$ is ergodic. Even more, $(Y_t)_{t \in \mathbb{Z}}$ is uniformly mixing with mixing coefficients satisfying $\phi(i) \leq const.(1-\kappa)^i$ for all $i \in \mathbb{N}$.*

The Proposition follows from known results for finite Markov chains, cf. Doukhan (1994, Th.1, Ch.2.4).

The geometrical decay of the mixing coefficients is typical for fixed, finite dimensional parametric models or for semiparametric models with a finite dimensional parametric part. It should not lead to a wrong conclusion that only very short range phenomena could be modelled with QVLMC's. Indeed, Theorem 2.1 in the next section discusses the broadness of the model class.

2.4 Range of QVLMC's

The class of stationary ergodic QVLMC's is broad enough to be weakly dense in the set of stationary processes on $(\mathbb{R}^d)^{\mathbb{Z}}$. Denote by

$$\pi_{t_1,\ldots,t_m} : (\mathbb{R}^d)^{\mathbb{Z}} \to (\mathbb{R}^d)^m, \; \pi_{t_1,\ldots,t_m}(y) = y_{t_1}, \ldots, y_{t_m} \; (t_1, \ldots, t_m \in \mathbb{Z}, \; m \in \mathbb{N})$$

the coordinate function and by '$\Rightarrow$' weak convergence.

Theorem 2.1 *Let P be a stationary process on $(\mathbb{R}^d)^{\mathbb{Z}}$ $(d \geq 1)$. Then, there exists a sequence $(P_n)_{n \in \mathbb{N}}$ of stationary, ergodic, $\mathbb{R}^d$-valued QVLMC's, such that*

$$P_n \circ \pi^{-1}_{t_1,\ldots,t_m} \Rightarrow P \circ \pi^{-1}_{t_1,\ldots,t_m} \text{ for all } t_1, \ldots, t_m \in \mathbb{Z}, \text{ for all } m \in \mathbb{N}.$$

A proof is given in the Appendix. For smooth P, a coarse quantization in the QVLMC is expected to be 'good' yielding a less complex model and typically a better approximation of P.

2.5 Prediction with QVLMC's

In principle, any predictor can be computed in a QVLMC model. From now on, we focus only on the optimal mean squared error m-step ahead predictor $\mathbb{E}[Y_{n+m}|Y_{-\infty}^{n}]$ for Y_{n+m}, given the past values $Y_{-\infty}^{n}$ and its corresponding estimate. For a QVLMC it is easy to see that

$$\begin{aligned}&\mathbb{E}_{QVLMC}[Y_{n+m}|Y_{-\infty}^{n}] = \mathbb{E}_{QVLMC}[Y_{n+m}|c(X_{-\infty}^{n})]\\ = &\sum_{x_{n+1}^{n+m}\in\mathcal{X}^m} \mathbb{E}[Y_{n+m}|X_{n+m} = x_{n+m}] \prod_{j=0}^{m-1} P_c(x_{n+m-j}|c(x_{n+1}^{n+m-j-1}X_{-\infty}^{n})),\end{aligned} \tag{2.8}$$

where $x_{n+1}^{n+m-j-1}X_{-\infty}^{n} = x_{n+m-j-1}, x_{n+m-j-2}, \ldots, x_{n+1}, X_n, X_{n-1}, \ldots$ for $j \geq 1$ and $x_{n+1}^{n+m-j-1}X_{-\infty}^{n} = X_{-\infty}^{n}$ for $j = m-1$. The QVLMC thus models the conditional expectation as a function of (finitely many) past quantized values $X_{-\infty}^{n}$ rather than $Y_{-\infty}^{n}$. It is a remarkable property of the QVLMC model that it allows easily multi-step ahead predictions. This is in contrast to other nonlinear prediction techniques where multi-step forecasts can become intractable, for example for parametric SETAR and nonparametric AR models or also for the adaptive partitioning TSMARS scheme (Lewis and Stevens, 1991).

It is sometimes of interest to predict an instantaneous function of a future observation $g(Y_{n+m})$ or to estimate the conditional expectation $\mathbb{E}[g(Y_{n+m})|Y_{-\infty}^{n}]$ for a fixed $g : \mathbb{R}^d \to \mathbb{R}^q$ $(d, q \in \mathbb{N})$. As an example, consider prediction of the volatility $\mathbb{E}[Y_{n+m}^2|Y_{-\infty}^{n}] - \mathbb{E}[Y_{n+m}|Y_{-\infty}^{n}]^2$ in financial time series; in fact, for $m = 1$ the ARCH- and GARCH-models are describing such a conditional variance within a given parametric function class. For an overview about ARCH- and GARCH-models, cf. Shephard (1996). The QVLMC predictor is then

$$\begin{aligned}&\mathbb{E}_{QVLMC}[g(Y_{n+m})|Y_{-\infty}^{n}] = \mathbb{E}_{QVLMC}[g(Y_{n+m})|c(X_{-\infty}^{n})]\\ = &\sum_{x_{n+1}^{n+m}\in\mathcal{X}^m} \mathbb{E}[g(Y_{n+m})|X_{n+m} = x_{n+m}] \prod_{j=0}^{m-1} P_c(x_{n+m-j}|c(x_{n+1}^{n+m-j-1}X_{-\infty}^{n}))\end{aligned} \tag{2.9}$$

2.6 Interpretation as dynamic adaptive partitioning

For notational simplicity we consider here a stationary real-valued process $(Y_t)_{t\in\mathbb{Z}}$ and discuss in more details the issues (1) and (2) from section 1.

In general, a partition scheme for a stationary Markov process of order p models $\mathbb{E}[Y_t|Y_{-\infty}^{t-1}]$ as

$$\mathbb{E}_{partition}[Y_t|S_{t-1}] = \sum_{j=1}^{J} c_j 1_{[S_{t-1}\in R_j]} = \sum_{j=1}^{J} c_j 1_{[V_{t-1}=R_j]}, \tag{2.10}$$

with $S_{t-1} = Y_{t-p}^{t-1}$ the state vector, $\{R_j;\ j = 1, \ldots, J\}$ a partition of $\mathbb{R}^p$ and $(V_t)_{t\in\mathbb{Z}}$ the partition cell process defined by $V_{t-1} = \sum_{j=1}^{J} R_j 1_{[S_{t-1}\in R_j]}$. The coefficients c_j $(j = 1, \ldots, J)$ are constants, depending only on the index j of the partition element R_j. The model in (2.10) is thus a step function from $\mathbb{R}^p$ to $\mathbb{R}$ (we exclude here partitioning schemes like MARS which uses splines instead of step functions). For a static partitioning scheme, these coefficients are of the form

$$c_j = \mathbb{E}[Y_t | V_{t-1} = R_j] = m_p(R_j),\ m_p : \mathcal{B}^p \to \mathbb{R} \tag{2.11}$$

with $\mathcal{B}^p$ the Borel σ-algebra of $\mathbb{R}^p$. The fact that $(V_t)_{t\in\mathbb{Z}}$ is a stochastic process (where information form the past could be non-trivial) is *not* used with such general static partitioning.

Dynamic partitioning as proposed here with QVLMC's corresponds to formula (2.10) as follows, compare with formula (2.8). The dimension p of the state vector S_{t-1} is the order of the VLMC $(X_t)_{t\in\mathbb{Z}}$, $J = |\tau_c|$ is the size of the minimal state space of the VLMC $(X_t)_{t\in\mathbb{Z}}$ and the partition elements R_j are given by

$$R_j = \{y_1^p \in \mathbb{R}^p;\ c(x_1^p) = w_j \in \tau_c\},\ x_t = q(y_t),\ j = 1, \ldots, J.$$

The coefficients are

$$c_j = \sum_{x\in\mathcal{X}} \mathbb{E}[Y_t | X_t = x] \mathbb{P}_{VLMC}[X_t = x | V_{t-1} = R_j]$$

with $\mathbb{P}_{VLMC}[X_t = x | V_{t-1} = R_j] = \mathbb{P}[X_t = x | c(X_{t-p}^{t-1}) = w_j]$ $(w_j \in \tau_c)$ the probability induced by the VLMC $(X_t)_{t\in\mathbb{Z}}$. To make the correspondence to (2.11) more explicit, this can also be written as

$$\begin{aligned} c_j &= \sum_{v_t \in \{R_j;\ j=1,\ldots,J\}} m_1(v_t) \mathbb{P}_{VLMC}[V_t = v_t | V_{t-1} = R_j], \\ m_1(v_t) &= m_1(v_{1,t}) = \mathbb{E}[Y_t | Y_t \in v_{1,t}],\ v_t = v_{1,t} \times \ldots \times v_{p,t},\ m_1 : \mathcal{B}^1 \to \mathbb{R} \end{aligned} \tag{2.12}$$

with $\mathcal{B}^1$ the Borel σ-algebra of $\mathbb{R}^1$ and $\mathbb{P}_{VLMC}[V_t = v_t | V_{t-1} = v_{t-1}]$ the probability induced by the VLMC $(X_t)_{t\in\mathbb{Z}}$, i.e., $\mathbb{P}_{VLMC}[V_t = v_t | V_{t-1} = v_{t-1}] = \mathbb{P}[X_t = x_t | c(X_{t-p}^{t-1}) = w_j]$ with $x_t,\ w_j$ such that $q(y_t) = x_t$ for all $y_t \in v_{1,t}$ and $c((q(y_s))_{s=t-p}^{t-1}) = w_j$ for all $y_{t-p}^{t-1} \in v_{t-1}$. Dynamic partitioning essentially differs from static partitioning in the model for the coefficients c_j $(j = 1, \ldots, J)$. As described by (2.12), our dynamic partitioning models $(V_t)_{t\in\mathbb{Z}}$ as a Markov chain and uses a nonparametric function $m_1(.)$ with domain $\mathcal{B}^1$, involving only a one-dimensional structure. This is in contrast to static partitioning as described in (2.11) where no dynamic model for $(V_t)_{t\in\mathbb{Z}}$ is assumed and a nonparametric function $m_p(.)$ with p-dimensional domain $\mathcal{B}^p$ is used.

When estimating the coefficients c_j $(j = 1, \ldots, J)$ we then observe the following. Static partitioning is exposed to the curse of dimensionality with the nonparametric function $m_p(.)$: the partitions should be derived in a clever *data-adaptive* way in order to escape as much as possible from the curse of dimensionality, cf. Breiman et al. (1984) or Gersho and Gray (1992). In our approach here, dynamic partitioning is exposed to the curse of dimensionality with the model for the quantized process $(X_t)_{t\in\mathbb{Z}}$, and *not* because of the nonparametric functions $m_1(.)$. The parsimonious VLMC model for $(X_t)_{t\in\mathbb{Z}}$, yielding a Markov model for the partition cell process, is a good way to escape from the curse of dimensionality, cf. Rissanen (1983), Weinberger at al. (1995) or Bühlmann & Wyner (1997).

For dynamic partitioning of a stationary Markov process as proposed here we conclude the following.

(1) The geometry of the partition $\{R_j;\ j = 1, \ldots, J\}$ is often not so important for estimation of the nonparametric function $m_1(.)$ with one-dimensional domain $\mathcal{B}^1$, see formula (2.12). Quantization of single variables as proposed here can then be viewed as some sort of *canonical operation* for dynamic partitioning.

(2) The model for the partition cell process $(V_t)_{t\in\mathbb{Z}}$ should be parsimonious. A natural parameterization is given by a VLMC for the quantized process $(X_t)_{t\in\mathbb{Z}}$.

Accepting the informal statements (1) and (2) then says that the QVLMC model is with the above interpretation a *natural* and kind of *canonical* model for dynamic partitioning. Estimation of a QVLMC in section 3 also shows that such a dynamic partitioning scheme is adaptive, mainly in terms of finding adaptively a VLMC model for $(X_t)_{t\in\mathbb{Z}}$ but also with a data-driven quantization. Last, fitting a QVLMC is a *white box* mechanism with well interpretable dynamic structure in terms of a context tree (see Definition 2.3) and a one-dimensional (and thus simple) nonparametric structure for $\mathbb{E}[Y_t|X_t]$, for example in terms of a scatter plot.

2.7 Equivalence to quantized state VLMC's

Consider a stationary $\mathbb{R}^d$-valued process $(Y_t)_{t\in\mathbb{Z}}$ whose memory is (of variable length as) a function of the quantized values of the past only, that is $\mathbb{P}[Y_1 \le y_1|Y_{-\infty}^0] = \mathbb{P}[Y_1 \le y_1|X_{-\infty}^0]$ for all $y_1 \in \mathbb{R}^d$, where $X_s = q(Y_s)$ with $q(.)$ as in (2.1). Assume that the conditional distributions have densities $f(y_1|x_{-\infty}^0)$ with respect to Lebesgue. Analogously to the Definitions 2.1 and 2.2 for an $\mathcal{X}$-valued VLMC we define

$$\begin{aligned}
&\tilde{c}(x_{-\infty}^0) = x_{-\tilde{\ell}+1}^0,\\
&\tilde{\ell} = \min\{k;\ f(y_1|x_{-\infty}^0) = f(y_1|x_{-k+1}^0) \text{ for Lebesgue almost all } y_1 \in \mathbb{R}^d\}\\
&(\tilde{\ell} \equiv 0 \text{ corresponds to independence}). \qquad (2.13)
\end{aligned}$$

We say that $(Y_t)_{t\in\mathbb{Z}}$ is a stationary quantized state variable length Markov chain (QSVLMC) of order p if $\max_{x_{-\infty}^0} |\tilde{c}(x_{-\infty}^0)| = p$.

The notion of a QSVLMC is more general than of a QVLMC from section 2.2. In connection with partitioning, which always reduces the state space of a Markov chain to a finite structure, the QSVLMC is in some sense even the most general. We argue now that it needs only little additional restrictions so that a QSVLMC is equivalent to the simpler QVLMC. Let us assume the following.

(B) The conditional densities of the stationary QSVLMC satisfy $f(y_1|x_{-\infty}^1) = f(y_1|x_1)$ for Lebesgue almost all $y_1 \in \mathbb{R}^d$ and all $x_{-\infty}^0 \in \mathcal{X}^\infty$ with $x_s = q(y_s)$ $(s \in \mathbb{Z})$.

This assumption (B) says that the variable Y_t given the instantaneous quantized information $X_t = q(Y_t)$ is independent from the past. A QVLMC satisfies (B).

Theorem 2.2 *Let $(Y_t)_{t\in\mathbb{Z}}$ be a QSVLMC with context function $\tilde{c}(.)$. Then the following holds.*

(i) The quantized process $(X_t)_{t\in\mathbb{Z}}$ with $X_t = q(Y_t)$ is a VLMC with context function $c(.)$ such that

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

(ii) If in addition assumption (B) holds, then $(Y_t)_{t\in\mathbb{Z}}$ is a QVLMC with $c(.) = \tilde{c}(.)$.

A proof is given in the Appendix. Theorem 2.2 explains that when insisting on assumption (B), a QVLMC is as general as a QSVLMC and hence in some sense the most general in the set-up of partitioning schemes.

3 Fitting of QVLMC's

We first have to find an appropriate quantizer q. In the univariate case, a practical procedure of choosing q when $N = |\mathcal{X}| \geq 2$ is specified is given by the sample quantiles $\hat{F}^{-1}(.)$ of the data,

$$\hat{q}(y) = \begin{cases} 0, & \text{if } -\infty < y \leq \hat{F}^{-1}(1/N) \\ x, & \text{if } \hat{F}^{-1}(x/N) < y \leq \hat{F}^{-1}(\frac{x+1}{N}) \ (x = 1, \ldots, N-2) \ , \\ N-1, & \text{if } \hat{F}^{-1}(\frac{N-1}{N}) < y < \infty. \end{cases} \tag{3.1}$$

yielding an interval partition with equal number of observations per partition cell. In the multivariate case, we could use a quantizer as in (2.6) with q_j estimated as in (3.1) in terms of the quantiles of the j-th individual series. The choice of an appropriate q or an appropriate size N of $\mathcal{X}$ could be given by the application. As examples we mention extreme events in engineering or finance with $N = 2$ for coding extreme or $N = 3$ for coding lower- and upper-extreme. Or, the choice of an appropriate q can be viewed as a model selection problem, this is discussed in section 5. We assume in the rest of this section that q is given and proceed as it would be correct. Given data $Y_1, \ldots, Y_n$ from a QVLMC, it then remains to estimate the cell densities $\{f_x(.);\ x = 0, 1, \ldots, N-1\}$ and the probability structure of the quantized $(X_t)_{t\in\mathbb{Z}}$, modelled as a VLMC.

3.1 Context algorithm

Given data $X_1, \ldots, X_n$ from a VLMC P_c on $\mathcal{X}^{\mathbb{Z}}$ (assuming that q is the correct quantizer), the aim is to find the underlying context function $c(.)$ and an estimate of P_c. (We identify the VLMC $(X_t)_{t\in\mathbb{Z}}$ with its probability measure P_c). In the sequel we always make the convention that quantities involving time indices $t \notin \{1, \ldots, n\}$ equal zero (or are irrelevant). Let

$$N(w) = \sum_{t=1}^{n} 1_{[X_t^{t+|w|-1} = w]},\ w \in \mathcal{X}^{|w|}, \tag{3.2}$$

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

$$\hat{P}(w) = N(w)/n,\ \hat{P}(x|w) = \frac{N(xw)}{N(w)},\ xw = (x_{|x|}, \ldots, x_2, x_1, w_{|w|}, \ldots, w_2, w_1). \tag{3.3}$$

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

$$\Delta_{wu} = \sum_{x\in\mathcal{X}} \hat{P}(x|wu) \log(\frac{\hat{P}(x|wu)}{\hat{P}(x|w)}) N(wu) \geq K \text{ for all } wu \in \hat{\tau}^T \ \ (u \in \mathcal{X}) \tag{3.4}$$

with $K = K_n \sim C\log(n),\ C > 2|\mathcal{X}| + 3$ a cut-off to be chosen by the user.

Step 1 Given $\mathcal{X}$-valued data $X_1, \ldots, X_n$, fit a maximal context tree, i.e., search for the context function $c_{max}(.)$ with terminal node context tree representation τ^T_{max} (see Definition 2.3), 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:

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

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

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

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

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). Prune $wu = x^0_{-\ell+1}$ to $w = x^0_{-\ell+2}$ if

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

with $K = K_n \sim C \log(n)$, $C > 2|\mathcal{X}|+3$ and $\hat{P}(.|.)$ as defined in (3.3). Decision about pruning for every terminal node in $\tau^T_{(0)}$ yields a (possibly) smaller tree $\tau_{(1)} \preceq \tau^T_{(0)}$. Construct the terminal node context tree $\tau^T_{(1)}$.

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

Step 4 If interested in probability sources, estimate the transition probabilities $P_c(x_1|c(x^0_{-\infty}))$ by $\hat{P}(x_1|\hat{c}(x^0_{-\infty}))$, where $\hat{P}(.|.)$ is defined as in (3.3).

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 tree are considered, in a hierarchical way, to be less relevant.

Consistency for finding the underlying true context function $c(.)$ and the VLMC probability distribution P_c goes back to Weinberger et al. (1995). For the algorithm described here, consistency even in an asymptotically infinite dimensional setting has been given in Bühlmann & Wyner (1997), where also more detailed descriptions of the context algorithm and cross-connections can be found. For deriving these results, we need some technical assumptions which we state in section 4.

The estimated minimal state space $\tau_{\hat{c}}$ from the context algorithm determines a data-adaptive partition of the space of past values $\{(x_s)^{t-1}_{-\infty};\ (x_s)^{t-1}_{-\infty} \in \mathcal{X}^{\infty}\}$ for the variable

X_t. The partitioning scheme is *non-recursive* and the tree growing architecture, namely the construction of the maximal tree τ_{max} in Step 1 of the context algorithm, covers very many imaginable models for categorical variables. This is in contrast to recursive tree growing procedures which sometimes exclude interesting sub-models due to the recursive partitioning. For an overview see Gersho & Gray (1992).

The estimation of the minimal state space τ_c is done solely on the basis of the quantized data $X_1, \ldots, X_n$. The question is if equivalence to fitting with $\mathbb{R}^d$-valued data holds. Let us assume the following.

(C) Estimation of the minimal state space τ_c of the QVLMC (or the underlying VLMC) is exclusively based on (possibly multiple) use of the log-likelihood ratio statistic

$$\tilde{\Delta}_{\tau_{c_1},\tau_{c_2}}(Y_1^n) = \log(\frac{\hat{f}_{\tau_{c_1}}(Y_1^n)}{\hat{f}_{\tau_{c_2}}(Y_1^n)}),$$

where τ_{c_1}, τ_{c_2} are (possibly different pairs of) context trees as in Definition 2.3 and

$$\log(\hat{f}_{\tau_{c_i}}(Y_1^n)) = \sum_{t=p+1}^{n} \log(\hat{f}(Y_t|c_i(X_{t-p}^{t-1}))) \ (i = 1, 2)$$

is the log-likelihood of an estimated QVLMC model with $c_i(.)$ induced by $\tau = \tau_{c_i}$ $(i = 1, 2)$, p the maximal order of $c_1(.)$ and $c_2(.)$, and $\hat{f}(Y_t|c_i(X_{t-p}^{t-1})) = \hat{f}_{X_t}(Y_t)\hat{P}(X_t|c_i(X_{t-p}^{t-1}))$ an estimate in the QVLMC model for $f(Y_t|c_i(X_{t-p}^{t-1})) = f_{X_t}(Y_t)P_{c_i}(X_t|c_i(X_{t-p}^{t-1}))$ $(i = 1, 2)$ with $\hat{f}_x(.)$ arbitrary and $\hat{P}(.|.)$ as in (3.3).

Assumption (C) is quite natural in the set-up of model selection. Neglecting the (minor) effect of different orders p for different τ_{c_i}'s, the context algorithm in section 3.1 with any estimator for $f_x(.)$ satisfies (C).

Proposition 3.1 *Assume that $(Y_t)_{t\in\mathbb{Z}}$ is a QVLMC with minimal state space τ_c. Then, any estimate of τ_c satisfying (C) is solely based on the quantized data $X_1, \ldots, X_n$.*

A proof is given in the Appendix. Proposition 3.1 also justifies to use the context algorithm as given in section 3.1 for estimation of the minimal state space τ_c. This is in contrast to many static partitioning schemes where the predictor variables are used in a quantized form but the response variable goes into the fitting of a partition as an $\mathbb{R}^d$-valued variable. As an example, we mention CART (Breiman et al., 1984).

3.2 Estimation of cell densities and cumulative probabilities

The cell densities $\{f_x(.); x \in \mathcal{X}\}$ can be estimated by some smoothing technique, e.g., a kernel estimator

$$\hat{f}_x(y) = \frac{n^{-1}h^{-d}\sum_{t=1}^{n} K(\frac{y-Y_t}{h})1_{[X_t=x]}}{n^{-1}N(x)}, \ y \in \mathbb{R}^d \tag{3.5}$$

with $N(x)$ as in (3.2), $K(.)$ a probability density function in $\mathbb{R}^d$ and h a bandwidth with $h = h(n) \to 0$ and typically $nh^{d+4} \to C$ $(n \to \infty)$ with $0 < C < \infty$ a constant. See for example Silverman (1986, Ch.4). Asymptotic properties of the estimates are given in section 4.

Instead of the densities $f_x(.)$, the cumulative probabilities of the observations are often of more interest. Then, one can use directly empirical distribution functions which makes smoothing, and thus selection of a bandwidth h, unnecessary. We estimate $\mathbb{P}[Y_t \in E|X_t = x]$ and $\mathbb{P}[Y_t \in E]$ for some (measurable) set E by

$$\hat{\mathbb{P}}[Y_t \in E|X_t = x] = \frac{n^{-1}\sum_{t=1}^{n} 1_{[Y_t \in E]} 1_{[X_t = x]}}{n^{-1}N(x)}, \tag{3.6}$$

$$\hat{\mathbb{P}}[Y_t \in E] = n^{-1}\sum_{t=1}^{n} 1_{[Y_t \in E]} = \sum_{x \in \mathcal{X}} \hat{\mathbb{P}}[Y_t \in E|X_t = x] n^{-1}N(x). \tag{3.7}$$

Asymptotic properties of these estimates are also discussed in section 4.

3.3 Estimated predictors

For the theoretical predictor in (2.8), we estimate $\mathbb{E}[Y_t|X_t = x]$ with $\overline{Y}_x = N(x)^{-1}\sum_{t=1}^{n} Y_t 1_{[X_t = x]}$, and using the estimated context function and transition probabilities for the VLMC $(X_t)_{t\in\mathbb{Z}}$ from Step 4 in section 3.1, we construct the plug-in estimator

$$\hat{Y}_{n+m} = \hat{\mathbb{E}}_{QVLMC}[Y_{n+m}|Y_1^n] = \sum_{x_{n+1}^{n+m} \in \mathcal{X}^m} \overline{Y}_{x_{n+m}} \prod_{j=0}^{m-1} \hat{P}(x_{n+m-j}|\hat{c}(x_{n+1}^{n+m-j-1}X_1^n)) \tag{3.8}$$

(see formula (2.8) for a proper definition of $x_{n+1}^{n+m-j-1}X_1^n$).
We use here the notation $\hat{Y}_{n+m}$ exclusively for the predictor which is estimated from the data. Asymptotic properties of $\hat{Y}_{n+m}$ are given in section 4.

The estimation of the predictor in (2.9) is constructed analogously,

$$\hat{\mathbb{E}}_{QVLMC}[g(Y_{n+m})|Y_1^n] = \sum_{x_{n+1}^{n+m} \in \mathcal{X}^m} \overline{g(Y)}_{x_{n+m}} \prod_{j=0}^{m-1} \hat{P}(x_{n+m-j}|\hat{c}(x_{n+1}^{n+m-j-1}X_1^n)), \tag{3.9}$$

with $\overline{g(Y)}_x = N(x)^{-1}\sum_{t=1}^{n} g(Y_t)1_{[X_t = x]}$.

4 Asymptotic inference

In order to analyze the asymptotic behavior of the estimates in section 3 we tacitly assume that the involved quantizer q is correct and make the following assumptions for the data generating process $(Y_t)_{t\in\mathbb{Z}}$.

(D1) $(Y_t)_{t\in\mathbb{Z}}$ is an $\mathbb{R}^d$-valued QVLMC satisfying condition (A) from section 2.3 and $\min_{w\in\tau_c} P_c(w) > 0$, $\min_{wu\in\tau_c, u\in\mathcal{X}} \sum_{x\in\mathcal{X}} |P_c(x|wu) - P_c(x|w)| > 0$, $\min_{x\in\mathcal{X}, w\in\tau_c} P_c(x|w) > 0$.

(D2) For all $x \in \mathcal{X}$ corresponding to the $\mathbb{R}^d$-valued QVLMC, $\int_{\mathbb{R}^d} \|z\|^2 f_x(z)dz < \infty$.

Condition (D1) is a regularity condition for the true underlying VLMC $(X_t)_{t\in\mathbb{Z}}$ which is also employed in Bühlmann & Wyner (1997). Condition (D2) implies $\mathbb{E}\|Y_t\|^2 < \infty$.

Assuming (D1), we then can show,

$$\mathbb{P}[\hat{c}(.) = c(.)] = o(1)\ (n \to \infty). \tag{4.1}$$

Since the event $\hat{c}(.) \neq c(.)$ has probability tending to zero, the estimated transition probabilities $\hat{P}(x_1|\hat{c}(x_{-\infty}^0))$ behave asymptotically like the maximum likelihood estimator $P(x_1|c(x_{-\infty}^0))$ when the context function $c(.)$ would be known. Thus under the assumptions in (D1), for every $x_{-\infty}^1 \in \mathcal{X}^\infty$,

$$n^{1/2}(\hat{P}(x_1|\hat{c}(x_{-\infty}^0)) - P(x_1|w)) \Rightarrow \mathcal{N}(0, (I^{-1})_{x_1 w, x_1 w})\ (n \to \infty),\ w = c(x_{-\infty}^0), \tag{4.2}$$

where I^{-1} is the inverse Fisher information matrix, whose values are given in the Appendix, formula (A.7).

For the cumulative probabilities as in (3.6) and (3.7) we obtain, assuming (D1),

$$n^{1/2}(\hat{\mathbb{P}}[Y_t \in E|X_t = x] - \mathbb{P}[Y_t \in E|X_t = x]) \Rightarrow \mathcal{N}(0, \nu_{as}^2(E, x))\ (n \to \infty), \tag{4.3}$$
$$n^{1/2}(\hat{\mathbb{P}}[Y_t \in E] - \mathbb{P}[Y_t \in E]) \Rightarrow \mathcal{N}(0, \xi_{as}^2(E))\ (n \to \infty). \tag{4.4}$$

A description of the values $\nu_{as}^2(E, x)$ and $\xi_{as}^2(E)$ is given in the Appendix, formulae (A.8) and (A.9). From the formulae (4.2) and (4.3) we also can state the asymptotic behavior of the m step ahead predictor in (3.8). Assuming (D1) and (D2), for y_1^n fixed and m finite,

$$n^{1/2}(\hat{\mathbb{E}}_{QVLMC}[Y_{n+m}|Y_1^n = y_1^n] - \mathbb{E}_{QVLMC}[Y_{n+m}|s]) \Rightarrow \mathcal{N}_d(0, \zeta_{as}^2(m, s))\ (n \to \infty), \tag{4.5}$$

where $s = c(q(y_{n-p+1}^n))$ is the state of the VLMC at time n and p the order of the VLMC. A discussion of the value $\zeta_{as}^2(m, s)$ and justification of the formulae (4.1)-(4.5) are given in the Appendix.

Finally, we can also obtain results for the estimated cell densities. We assume that for all $x \in \mathcal{X}$, $f_x(.)$ is bounded and twice continuously differentiable. The kernel $K(.)$ is a probability density in $\mathbb{R}^d$ with compact support, bounded and continuously differentiable, $\int \|z\|^2 K(z)dz < \infty$, $\int z_j K(z)dz = 0$ $(j = 1, \ldots, d)$ and $\|z\|^d K(z) \to 0$ $(\|z\| \to \infty)$. The bandwidth h is of the order $n^{-1/(d+4)}$, balancing asymptotic bias $B_{as}(x, y)$ and variance $\sigma_{as}^2(x, y)$. Assuming in addition (D1), we obtain for all $x \in \mathcal{X}$ and $y \in \overset{\circ}{I}_x$ (y an interior point of I_x),

$$(nh^d)^{1/2}(\hat{f}_x(y) - f_x(y)) \Rightarrow \mathcal{N}(B_{as}(x, y), \sigma_{as}^2(x, y))\ (n \to \infty), \tag{4.6}$$

A precise description of the values $B_{as}(x, y)$ and $\sigma_{as}^2(x, y)$ is given in the Appendix, formula (A.11). The assumption about the kernel $K(.)$ and the boundedness of the densities $f_x(.)$ can be modified, cf. Boente & Fraiman (1995).

The expressions for the asymptotic variances $\nu_{as}^2(E, x)$, $\xi_{as}^2(E)$ and $\zeta_{as}^2(m, s)$ are quite complicated and particularly for the latter almost impossible to estimate analytically. The block bootstrap (Künsch, 1989) is here a powerful machinery to estimate the limiting distributions of the discussed quantities, say for the sake of constructing confidence regions.

Asymptotic inference is given here if the data is a finite realization of a QVLMC satisfying (D1) and (D2) and if the quantizer q (viewed as a nuisance parameter) is correct. The latter is of course quite a hypothetical assumption: we aim to do further research on asymptotic theory when considering the quantizer q as an estimate from the data.

5 Model selection

Choosing a quantizer $q : \mathbb{R}^d \to \mathcal{X}$ and a minimal state space τ_c (or equivalently the context function $c(.)$) is here proposed in a data-driven fashion. For obtaining simplicity and manageability in selecting a model we assume a Gaussian component quasi-likelihood structure which allows us then to develop a strategy in a parametric set-up, although the original problem is of semi- or nonparametric nature.

We focus first on the univariate case. The log-likelihood function of a QVLMC (conditional on the first p observations) is

$$\ell(Y_1, \ldots, Y_n) = \sum_{t=p+1}^{n} \log(f_{X_t}(Y_t)) + \sum_{t=p+1}^{n} \log(P_c(X_t | c(X_{t-p}^{t-1})),$$

where p is the order of the underlying VLMC. Denote by $\mu_x = \mathbb{E}[Y_t | X_t = x]$ and $\sigma_x^2 = \mathrm{Var}(Y_t | X_t = x)$. By assuming $f_x(y) = (2\pi\sigma_x^2)^{-1/2} \exp(-(y - \mu_x)^2 / (2\sigma_x^2))$ (although $supp(f_x) = \mathbb{R}$) we then consider the Gaussian component quasi-log-likelihood function

$$\ell_{quasi}(\theta; Y_1, \ldots, Y_n) = \sum_{t=p+1}^{n} \log\{(2\pi\sigma_{X_t}^2)^{-1/2} \exp(-(Y_t - \mu_{X_t})^2 / (2\sigma_{X_t}^2))\}$$
$$+ \sum_{t=p+1}^{n} \log(P_c(X_t | c(X_{t-p}^{t-1})), \tag{5.1}$$

where $\theta = (\mu_0, \ldots, \mu_{N-1}, \pi)$ with $\pi_{wx} = \mathbb{P}[X_t = x | c(X_{t-p}^{t-1}) = w]$, $w \in \tau_c$, $x \in \{0, \ldots, N-2\}$. The maximum quasi-likelihood estimator (MQLE)

$$\hat{\theta}_{MQLE} = \mathrm{argmin}_{\theta}(-\ell_{quasi}(\theta; Y_1, \ldots, Y_n) \tag{5.2}$$

yields then the parameter values which are used for the estimated predictor in (3.8). For the prediction problem we thus can restrict our attention to the quasi-likelihood function in (5.1) and the MQLE estimator in (5.2). The quasi-likelihood function itself is not meant to describe the whole underlying distribution of the observations but rather the characteristics of the conditional expectation $\mathbb{E}[Y_t | Y_1^{t-1}]$, cf. McCullagh & Nelder (1989, Ch.9). In the parametric case, in which we are now due to the Gaussian component quasi-likelihood assumption, it is known that a proper AIC-type criterion is of the form

$$-2\ell_{quasi}(\hat{\theta}; Y_1, \ldots, Y_n) + \text{penalty-term}.$$

The penalty term is generally of the form $2tr|IJ^{-1}|$ where I is the asymptotic variance of the score statistic and J the expectation of the Hessian of the log-likelihood at the observations Y_1^n, cf. Shibata (1989) or Bühlmann (1997). Generally, it is quite complicated to estimate $tr|IJ^{-1}|$ from the data. For the simpler case of an $\mathcal{X}$-valued VLMC alone (with no quantization involved) a bootstrap approach has been proposed in Bühlmann (1997) which is computationally already quite expensive. To obtain a manageable criterion we further assume that the difference between I and J is negligible, since $I = J$ if the model under consideration is the correct one. With such a commonly made assumption, the penalty term becomes $2\dim(\theta)$. In our case, $\dim(\theta) = N + |\tau_c|(N-1)$ $(N = |\mathcal{X}|)$. By replacing σ_x^2 with $\hat{\sigma}_x^2 = (N(x)-1)^{-1}\sum_{t=1}^n (Y_t - \overline{Y}_x)^2 1_{[X_t=x]}$, our model selection criterion then becomes

$$\begin{aligned} M^2 &= -2\ell_{quasi}(\hat{\theta}; Y_1, \ldots, Y_n) + 2(N + |\tau_c|(N-1)) \\ &= \sum_{t=p+1}^{n} \{(Y_t - \overline{Y}_{X_t})^2/\hat{\sigma}^2_{X_t} + \log(2\pi\hat{\sigma}^2_{X_t})\} \\ &- 2\sum_{t=p+1}^{n} \log(\hat{P}(X_t|c(X_{t-p}^{t-1}))) + 2(N + |\tau_c|(N-1)), \end{aligned}$$

where $\hat{P}(.|.)$ is given in (3.3). Note that the quantizer q enters implicitly. Theoretically we would then search for the model, i.e, the quantizer q and the context function $c(.) = c_q(.)$ (depending on q) which minimize M^2. However, the search over all context functions is infeasible. A remedy proposed in Bühlmann (1997) is to search for an optimal cut-off parameter K in the context algorithm, see Step 2 in section 3.1. Then, the optimal model, or now the optimal tuning of the algorithm, is specified by the quantizer q and the cut-off parameter K which minimize

$$\begin{aligned} M^2(q,K) &= \sum_{t=p+1}^{n} \{(Y_t - \overline{Y}_{X_t})^2/\hat{\sigma}^2_{X_t} + \log(2\pi\hat{\sigma}^2_{X_t})\} \\ &- 2\sum_{t=p+1}^{n} \log(\hat{P}(X_t|\hat{c}_K(X_{t-p}^{t-1}))) + 2(N + |\tau_{\hat{c}_K}|(N-1)). \end{aligned}$$

where $\hat{c}_K$ is the estimated context function for the $\mathcal{X}$-valued VLMC (depending on K) and $\hat{P}(.|.)$ as in (3.3). Note that for given q, the search for an optimal cut-off K is only affected by the term and $-2\sum_{t=p+1}^n \log(\hat{P}(X_t|\hat{c}_K(X_{t-p}^{t-1})))$, thus being exactly the same as when tuning the context algorithm for categorical valued VLMC, see Bühlmann (1997).

For the multivariate QVLMC model, we propose to choose the quantizer q and the optimal cut-off K of the context algorithm as the minimizer of

$$\begin{aligned} M_d^2(q,K) &= \sum_{t=p+1}^{n} \{(Y_t - \overline{Y}_{X_t})'\hat{\Sigma}^{-1}_{X_t}(Y_t - \overline{Y}_{X_t}) + d\log(2\pi) + \log(|\hat{\Sigma}_{X_t}|)\} \\ &- 2\sum_{t=p+1}^{n} \log(\hat{P}(X_t|\hat{c}_K(X_{t-p}^{t-1}))) + 2(N + |\tau_{\hat{c}_K}|(N-1)) \end{aligned}$$

where $\hat{\Sigma}_x = (N(x)-1)^{-1}\sum_{t=1}^n (Y_t - \overline{Y}_x)(Y_t - \overline{Y}_x)' 1_{[X_t=x]}$ and $\hat{P}(.|.)$ as in (3.3).

6 Numerical examples

We study the predictive performance of the dynamic adaptive QVLMC scheme for simulated and real data by considering the one-step ahead predictor from (3.8). The sample size is denoted by n. We then compute a predictor for the next m observations: we do not re-estimate the predictor, it is always based on the first n observations. We calculate

$$\mathrm{PE} = n^{-1}\sum_{t=n+1}^{n+m} (\hat{Y}_t - Y_t)^2 \tag{6.1}$$

with $\hat{Y}_t$ the predictor estimated on the first variables $Y_1, \ldots, Y_n$ and evaluated on Y_1^{t-1}. By ergodicity and the mixing property, the quantity PE approximates (as the number m of predicted observations grows) the final prediction error conditioned on the observed data $\mathbb{E}[(\hat{Z}_t - Z_t)^2 | Y_1^n]$, where $(Z_t)_{t\in\mathbb{Z}}$ is an independent copy of $(Y_t)_{t\in\mathbb{Z}}$ and $\hat{Z}_t$ is the predictor estimated from the data Y_1^n and evaluated as a function of the past values $Z_{-\infty}^{t-1}$.

We compute the measure PE of actual predictive performance for the QVLMC predictor in (3.8) for various quantizers q and cut-off parameter K in the context algorithm in section 3.1 being always $\chi^2_{N-1;0.95}/2$ which is often a reasonable value, cf. Bühlmann & Wyner (1997) and Bühlmann (1997). In the following examples, the quantizer $q = \hat{q}$ is estimated from the data as in (3.1), unless specified otherwise. Varying over $\hat{q}$ then results in varying over $N = |\mathcal{X}|$. For comparison, we also compute PE for the predictor of an $AR(p)$ model with p chosen by the minimum AIC criterion in the range $10\log_{10}(n)$. In the case of the QVLMC predictor, we also give the measure $M^2(q, K)$ ($= M^2(N)$ by our choice of $q = \hat{q}$ and K) for model selection from section 5 as an estimate of predictive performance.

6.1 Simulated data

We consider various data-generating models representing different kinds of stationary processes. Most often, we consider $n = 4000$ (once $n = 500$) and we predict $m = 1000$ values.

Univariate QVLMC with $N = 4$. Let the true quantizer q be given by the partition in (2.2) with

$$I_0 = (-\infty, -0.396],\ I_1 = (-0.396, 0],\ I_2 = (0, 1.189],\ I_3 = (1.189, \infty).$$

The distribution of $Y_t | X_t = x \sim f_x(y)dy$ is given by

$$Y_t = \begin{cases} \min(Z_t, -0.396),\ Z_t \sim \mathcal{N}(-0.793, 0.039) & \text{if } X_t = 0, \\ \min(\max(Z_t, -0.395), 0),\ Z_t \sim \mathcal{N}(-0.198, 0.020) & \text{if } X_t = 1, \\ \min(\max(Z_t, 0.001), 1.189),\ Z_t \sim \mathcal{N}(0.594, 0.039) & \text{if } X_t = 2, \\ \max(Z_t, 1.190),\ Z_t \sim \mathcal{N}(1.585, 0.039) & \text{if } X_t = 3. \end{cases}$$

The underlying VLMC $(X_t)_{t\in\mathbb{Z}}$ is given by Figure 6.1 in terms of the context tree and transition probabilities $(P(0|w), \ldots, P(3|w))$ with w the terminal nodes in the tree. The QVLMC model is specified so that $\mathrm{Var}(Y_t) \approx 1$. The true model is not found by our restricted QVLMC scheme in this simulation study: the quantizer $\hat{q}$ with $N = 4$ as in (3.1) is not approaching the true q because of the location of the true quantiles of the one-dimensional marginal distribution.

Table 6.1 summarizes the results for $n = 4000$. We abbreviate with 'QVLMC, N' the QVLMC predictor with quantizer as in (3.1) determined by N; 'QVLMC, true q' is the QVLMC scheme with the true quantizer q; 'oracle' is the predictor with known probability distribution of the underlying data generating process. The model-dimension indicates the complexity of the scheme; for an AR(p) predictor, the model-dimension is $p + 1$ due to the fact that we apply a mean-correction first. The efficiency of the best selected QVLMC scheme according to $M^2(q, K)$ (which is also the best among the fully data-driven QVLMC cases in Table 6.1) relative to the oracle, defined as the ratio of the corresponding PE's,

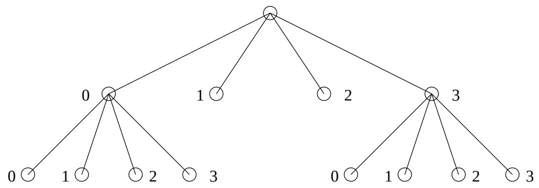

Figure 6.1: Context tree of QVLMC model.

method	model-dimension	PE	$M^2(q,K)$
QVLMC, $N = 16$	1861	0.506	7592.6
QVLMC, $N = 12$	1068	0.446	6625.2
QVLMC, $N = 9$	553	0.489	6839.4
QVLMC, $N = 6$	211	0.435	5564.2
QVLMC, $N = 4$	193	0.528	6552.7
QVLMC, true q	148	0.423	3425.0
AR	5	0.528	–
oracle	–	0.416	–

Table 6.1: Performances for QVLMC model, $n = 4000$.

is $0.435/0.416 = 1.05$ and thus close to optimal. For comparison, the efficiency of the best selected QVLMC scheme relative to the AR predictor is $0.435/0.528 = 0.82$ which is a substantial gain. Interestingly, the performance of the QVLMC scheme with known q having $N = 4$ is most competing with the QVLMC scheme with $\hat{q}$ as in (3.1) and $N = 6$, and not $N = 4$: as mentioned above, this can be explained with the fact that $\hat{q}$ in (3.1) with $N = 4$ is never approximating the true q.

Univariate TAR(1). Consider a threshold autoregressive model of order 1,

$$Y_t = \begin{cases} 0.9Y_{t-1} + Z_t & \text{if } Y_{t-1} \leq -1.143 \\ -0.9Y_{t-1} + Z_t & \text{if } Y_{t-1} > -1.143 \end{cases}$$

with $(Z_t)_{t\in\mathbb{Z}}$ an i.i.d. sequence, $Z_t \sim \mathcal{N}(0, 0.209)$ and Z_t independent from Y_s for all $s < t$. Again, the model is specified so that $\mathrm{Var}(Y_t) \approx 1$.

The results for $n = 4000$ (with the same notation as in Table 6.1) are given in Table 6.2. The efficiency of the best selected QVLMC scheme according to $M^2(q,K)$ (which is not, but almost the best among the QVLMC cases in Table 6.2) relative to the oracle is $0.256/0.209 = 1.22$. The efficiency of the best selected QVLMC scheme relative to the AR predictor is $0.256/0.728 = 0.35$ which is a huge gain.

method	model-dimension	PE	$M^2(q,K)$
QVLMC, $N=24$	576	0.251	7588.9
QVLMC, $N=20$	400	0.256	7352.2
QVLMC, $N=16$	256	0.261	7358.7
QVLMC, $N=12$	144	0.266	7415.8
QVLMC, $N=9$	97	0.327	7794.5
QVLMC, $N=6$	81	0.413	8480.6
AR	5	0.728	–
oracle	–	0.209	–

Table 6.2: Performances for TAR model, $n = 4000$.

method	model-dimension	PE	$M^2(q,K)$
QVLMC, $N=24$	599	0.666	9990.9
QVLMC, $N=20$	761	0.653	10095.1
QVLMC, $N=16$	796	0.642	9863.0
QVLMC, $N=12$	639	0.640	10036.6
QVLMC, $N=9$	321	0.671	10195.3
QVLMC, $N=6$	201	0.806	11436.5
AR	2	0.999	–
oracle	–	0.523	–

Table 6.3: Performances for nonparametric AR(2) in (6.2), $n = 4000$.

Univariate nonparametric AR(2). We consider here two types of general nonparametric AR(2) models, the first one being additive for the mean function in the lagged variables and with conditional heteroscedastic errors,

$$Y_t = 0.863\sin(4.636Y_{t-1}) + 0.431\cos(4.636Y_{t-2}) + (0.023 + 0.5Y_{t-1}^2)^{1/2}Z_t \quad (6.2)$$

with $(Z_t)_{t\in\mathbb{Z}}$ an i.i.d. sequence, $Z_t \sim \mathcal{N}(0,1)$ and Z_t independent from Y_s for all $s < t$. Again, the model is specified so that $\mathrm{Var}(Y_t) \approx 1$.

The results for $n = 4000$ (with the same notation as in Table 6.1) are given in Table 6.3. The efficiency of the best selected QVLMC scheme according to $M^2(q,K)$ relative to the oracle is $0.642/0.523 = 1.23$. The efficiency of the best selected QVLMC scheme relative to the AR predictor is $0.642/0.999 = 0.64$ which is a big gain.

The other nonparametric autoregressive model has an interaction term in the mean function,

$$Y_t = (0.5 + 0.9\exp(-2.354Y_{t-1}^2))Y_{t-1} - (0.8 - 1.8\exp(-2.354Y_{t-1}^2))Y_{t-2} + Z_t \quad (6.3)$$

with $(Z_t)_{t\in\mathbb{Z}}$ an i.i.d. sequence, $Z_t \sim \mathcal{N}(0, 0.425)$ and Z_t independent from Y_s for all $s < t$. Again, the model is specified so that $\mathrm{Var}(Y_t) \approx 1$. Such a model is also known as 'Exponential AR(2)'.

The results for $n = 4000$ and $n = 500$ (with the same notation as in Table 6.1) are given in Tables 6.4 and 6.5, respectively. The efficiencies of the best selected QVLMC scheme according to $M^2(q,K)$ relative to the oracle are $0.474/0.425 = 1.16$ ($n = 4000$) and $0.592/0.425 = 1.39$ ($n = 500$). The efficiencies of the best selected QVLMC scheme

method	model-dimension	PE	$M^2(q,K)$
QVLMC, $N = 24$	829	0.779	12381.3
QVLMC, $N = 20$	932	0.700	12056.1
QVLMC, $N = 16$	1126	0.558	11524.1
QVLMC, $N = 12$	870	0.482	10953.9
QVLMC, $N = 9$	489	0.474	10732.1
QVLMC, $N = 6$	196	0.521	10833.6
AR	10	0.842	–
oracle	–	0.425	–

Table 6.4: Performances for nonparametric AR(2) in (6.3), $n = 4000$.

method	model-dimension	PE	M^2
QVLMC, $N = 9$	89	0.805	1677.0
QVLMC, $N = 7$	109	0.650	1637.0
QVLMC, $N = 6$	81	0.584	1563.1
QVLMC, $N = 5$	57	0.592	1488.0
QVLMC, $N = 4$	40	0.601	1500.6
QVLMC, $N = 3$	27	0.646	1548.0
AR	7	0.868	–
oracle	–	0.425	–

Table 6.5: Performances for nonparametric AR(2) in (6.3), $n = 500$.

relative to the AR predictor are $0.474/0.842 = 0.56$ ($n = 4000$) and $0.592/0.868 = 0.68$ ($n = 500$) which are big gains. We see here, that even for the smaller sample size $n = 500$, the QVLMC scheme clearly outperforms the AR predictor.

Univariate AR(2). We consider here a Gaussian AR(2),

$$Y_t = 0.5Y_{t-1} - 0.8Y_{t-2} + Z_t$$

with $(Z_t)_{t\in\mathbb{Z}}$ an i.i.d. sequence, $Z_t \sim \mathcal{N}(0, 0.341)$ and Z_t independent from Y_s for all $s < t$. Again, the model is specified so that $\mathrm{Var}(Y_t) \approx 1$.

The results for $n = 4000$ are given in Table 6.6. We abbreviate with 'AR(2)' the predictor based on an AR(2) with the true order; otherwise, the notation is as in Table 6.1. The efficiency of the best selected QVLMC scheme according to $M^2(q,K)$ relative to the oracle is $0.430/0.341 = 1.26$. The efficiency of the best selected QVLMC scheme relative to the AR predictor is $0.430/0.356 = 1.21$ which is a moderate loss.

Bivariate nonparametric AR(1). We consider here a bivariate nonparametric AR(1),

$$\begin{aligned}
Y_{1,t} &= 1.107\sin(3.629Y_{1,t-1}) + 0.554\cos(3.598U_{t-1}) + (0.038 + 0.200U_{t-1}^2)^{1/2}Z_{1,t},\\
U_t &= 1.107\sin(3.598U_{t-1}) + 0.554\cos(3.629Y_{1,t-1}) + (0.038 + 0.200Y_{1,t-1}^2)^{1/2}Z_{2,t},\\
Y_{2,t} &= 4.721\left(\frac{\exp(U_t)}{1+\exp(U_t)} - 0.5\right)
\end{aligned}$$

with $(Z_{1,t})_{t\in\mathbb{Z}}$, $(Z_{2,t})_{t\in\mathbb{Z}}$ independent i.i.d. sequences, $Z_{1,t} \sim \mathcal{N}(0,1)$, $Z_{2,t} \sim \mathcal{N}(0,1)$ and $Z_{1,t}$, $Z_{2,t}$ independent from $Y_{1,s}, U_s$ for all $s < t$. The series $(U_t)_{t\in\mathbb{Z}}$ is only auxiliary for the

method	model-dimension	PE	$M^2(q,K)$
QVLMC, $N=24$	806	0.854	12901.5
QVLMC, $N=20$	1179	0.743	12761.0
QVLMC, $N=16$	1846	0.520	11779.4
QVLMC, $N=12$	1387	0.410	10628.4
QVLMC, $N=9$	657	0.407	10043.1
QVLMC, $N=6$	261	0.430	9940.5
AR	7	0.356	–
AR(2)	3	0.356	–
oracle	–	0.341	–

Table 6.6: Performances for AR(2), $n = 4000$.

method	model-dimension	$(\mathrm{PE}_1, \mathrm{PE}_2)$	PE_{tot}	$M_2^2(q,K)$
QVLMC, $(N_1 = N_2 = 5)$	625	$(0.594, 0.578)$	1.172	21048.6
QVLMC, $(N_1 = N_2 = 4)$	271	$(0.628, 0.596)$	1.224	20740.7
QVLMC, $(N_1 = N_2 = 3)$	249	$(0.960, 0.966)$	1.962	24223.1
AR	6	$(1.000, 0.996)$	1.996	–

Table 6.7: Performances for bivariate nonparametric AR(1), $n = 4000$.

definition of $(Y_{1,t}, Y_{2,t})_{t\in\mathbb{Z}}$. Again, the model is specified so that $\mathrm{Var}(Y_{1,t}) \approx 1,\ \mathrm{Var}(Y_{2,t}) \approx 1$.

The results for $n = 4000$ are given in Table 6.7. We abbreviate with 'QVLMC, $(N_1,\ N_2)$' the QVLMC predictor with quantizer as in (2.7) with q_j and corresponding values N_j $(j = 1, 2)$ as in (3.1) for the two individual series. We restrict here attention to the case $N_1 = N_2$ which is not a necessity. Denote by $\mathrm{PE}_j = n^{-1}\sum_{t=n+1}^{n+m}(\hat{Y}_{j,t} - Y_{j,t})^2$ $(j = 1, 2)$ and by $\mathrm{PE}_{tot} = \mathrm{PE}_1 + \mathrm{PE}_2$. Due to the non-Markovian character of $(Y_t)_{t\in\mathbb{Z}}$, the oracle PE is difficult to obtain. The minimum AIC AR approximation is given by a bivariate AR(1). The efficiency for PE_{tot} of the best selected QVLMC scheme according to $M_2^2(q,K)$ relative to the AR predictor is $1.224/1.996 = 0.61$ which is a big gain.

Summary. For the models studied here, the QVLMC scheme is substantially better than the linear AR technique, except for the AR(2) model where the loss against the AR predictor is not very big. Maybe most important in practice, the QVLMC is not very sensitive to the specification of the size N of the space $\mathcal{X}$, and the model selection criterion $M^2(q,K)$ works well.

6.2 Return-volume data from New York stock exchange

The data is about daily return of the Dow Jones index and daily volume of the New York Stock Exchange (NYSE) from the period July 1962 through June 1988, corresponding to 6430 days. * What we term 'volume' is the standardized aggregate turnover on the NYSE,

*This data is publicly available via the internet:
http://ssdc.ucsd.edu/ssdc/NYSE.Date.Day.Return.Volume.Vola.text

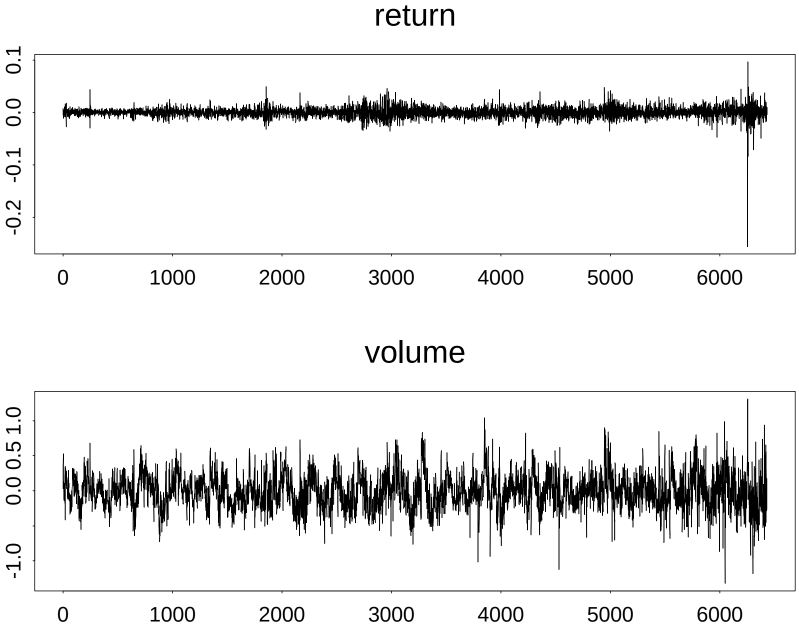

Figure 6.2: Daily return and volume.

$$\mathrm{Vol}_t = \log(F_t) - \log(\sum_{s=t-100}^{t-1} F_s/100),$$

where F_s equals the fraction of shares traded on day s. The return is defined as

$$\mathrm{Ret}_t = \log(D_t) - \log(D_{t-1}),$$

where D_t is the index of the Dow Jones on day t. Figure 6.2 shows these standardized time series; the time t=6253, where return $\mathrm{Ret}_t < -0.2$, corresponds to the 1987 crash. Besides this special structure around the crash, both series look stationary. The association between return and volume is an established fact, cf. Karpoff (1987).

We fit a bivariate QVLMC to the first $n = 6000$ observations. The quantizer q is as in (2.7) with q_j as in (3.1) for the two individual series.

In a first analysis we consider the actual predictive performance as in section 6.1, although there is probably little predictive information for the return series. We evaluate $m = 250$ one-step ahead predictors for the variables at times $6001, \ldots, 6250$, thus excluding the 1987 crash and the period afterwards. The results are given in Table 6.8. We use the same notation as in Table 6.7, PE_{tot} does not make much sense since the two series are on two very different scales. The best selected QVLMC model according to $M_2^2(d, K)$ uses $N_1 = N_2 = 5$. The differences by using other class sizes N_1, N_2 are not so big. Effective prediction of the return series is here not possible, the best selected QVLMC predictor is just as good as the global arithmetic mean; the AR predictor is even slightly worse than the global mean. It is questionable whether forecasting of these returns is even a possible task. Prediction of the volume series is successful: the efficiencies of the best selected QVLMC relative to the AR predictor and the global mean are $0.0583/0.0620 = 0.94$ and $0.0583/0.0783 = 0.74$, respectively.

method	model-dimension	$(\mathrm{PE}_{\mathrm{Ret}}, \mathrm{PE}_{\mathrm{Vol}})$	$M_2^2(q, K)$
QVLMC, $(N_1 = N_2 = 5)$	625	$(1.22\ \text{e-}04, 0.0583)$	-39347.3
QVLMC, $(N_1 = N_2 = 4)$	466	$(1.23\ \text{e-}04, 0.0600)$	-39044.1
QVLMC, $(N_1 = 3, N_2 = 9)$	755	$(1.23\ \text{e-}04, 0.0589)$	-39200.8
QVLMC, $(N_1 = 3, N_2 = 8)$	622	$(1.23\ \text{e-}04, 0.0594)$	-39315.7
QVLMC, $(N_1 = 9, N_2 = 3)$	703	$(1.23\ \text{e-}04, 0.0602)$	-38455.6
QVLMC, $(N_1 = 8, N_2 = 3)$	622	$(1.23\ \text{e-}04, 0.0605)$	-38750.6
AR	126	$(1.25\ \text{e-}04, 0.0620)$	–
global mean	1	$(1.22\ \text{e-}04, 0.0783)$	–

Table 6.8: Performances for return-volume data, estimation on $n = 6000$.

Regarding the return series, it is of interest to predict the volatility: given time t, we want to estimate $\mathrm{Var}(\mathrm{Ret}_{t+1}|\mathrm{Ret}_{-\infty}^{t}, \mathrm{Vol}_{-\infty}^{t-1})$. With the estimated bivariate QVLMC

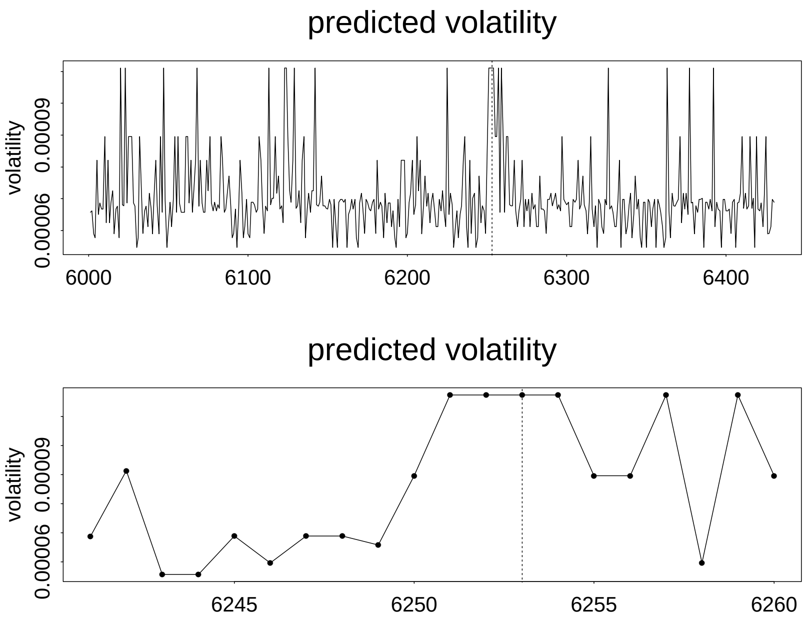

Figure 6.3: Predicted volatility $\hat{\mathrm{Var}}_{QVLMC}(\mathrm{Ret}_{t+1}|\mathrm{Ret}_1^t, \mathrm{Vol}_1^t)$. The vertical line indicates the 1987 crash point at $t = 6253$.

model with $N_1 = N_2 = 5$, based on the first 6000 observations, we computed

$$\begin{aligned} & \hat{\mathrm{Var}}_{QVLMC}(\mathrm{Ret}_{t+1}|\mathrm{Ret}_1^t, \mathrm{Vol}_1^t) \\ = \ & \hat{\mathbf{E}}_{QVLMC}[\mathrm{Ret}_{t+1}^2|\mathrm{Ret}_1^t, \mathrm{Vol}_1^t] - \hat{\mathbf{E}}_{QVLMC}[\mathrm{Ret}_{t+1}|\mathrm{Ret}_1^t, \mathrm{Vol}_1^t]^2, \ t = 6001, \ldots, 6430. \end{aligned}$$

given by formula (3.9). The result is plotted in Figure 6.3. There is a clear indication for a non-constant volatility, a feature which cannot be reflected with a multivariate ARMA model. A magnification around the 1987 crash shows that the one-step ahead predicted volatility was already high two days before the crash happened and is predicted to be high at the crash and its next day: among the time points $t = 6001, \dots, 6430$, the period around the crash is the longest with constantly high predicted volatility (4 days in a row). Note that because of using a partitioning scheme, there is only a finite number of (but still many) values for the predicted volatility $Var_{QVLMC}(\mathrm{Ret}_{t+1}|\mathrm{Ret}_{-\infty}^{t}, \mathrm{Vol}_{-\infty}^{t})$ $(t \in \mathbb{Z})$: extracting more information, in particular in the tails, is very difficult without making a strong parametric extrapolation.

7 Conclusions

The new notion of dynamic adaptive partitioning is here realized with fitting of QVLMC's. Of course, other approaches for dynamic adaptive partitioning are possible. However, the issues (1) and (2) from section 2.6 and the results in Theorem 2.2 and Proposition 3.1 indicate the special position of QVLMC's in dynamic adaptive partitioning. This is achieved by insisting on the mathematical, but rather natural, assumptions (B) and (C) in sections 2.7 and 3.1, respectively: weakening these assumptions could yield other dynamic adaptive partitioning schemes whose practical and theoretical potential has not been explored yet.

The range of asymptotic validity for the QVLMC class is larger than of adaptive (static) partitioning schemes like CART (Breiman et al., 1984) or MARS (Friedman, 1991) which are candidates for consistently estimating a high order nonparametric autoregressive model (for the conditional mean) but not directly for more complicated models, e.g., in the presence of conditional heteroscedastic errors. The QVLMC scheme allows also in a straightforward way the handling of multivariate data.

We do not believe that there is a uniformly best (or good) technique for the very broad framework of nonlinear stationary time series. The dynamic adaptive QVLMC scheme should be viewed as a potentially useful technique which is here justified by some theoretical arguments and some numerical results. It is an additional proposal being rather different from other known methodologies. In contrast to a black box mechanism, the components of the QVLMC scheme have a well interpretable structure.

Appendix

Proof of Theorem 2.1. For notational simplicity we give only the proof for the univariate case with $d = 1$. Let P be a stationary process on $\mathbb{R}^{\mathbb{Z}}$.

Step 1. Show that P can be approximated by a sequence of discrete, stationary distributions $(P_k)_{k \in \mathbb{N}}$ with P_k on $\Xi_k^{\mathbb{Z}}$, where Ξ_k is a finite space. By the Lebesgue decomposition Theorem, write $P = P_s + P_r$ with P_s discrete (singular with respect to Lebesgue) and P_r continuous. Trivially, P_s can be approximated by a sequence $(P_{s;k})_{k \in \mathbb{N}}$ with $P_{s;k}$ on $\Xi_{s;k}^{\mathbb{Z}}$, $\Xi_{s;k}$ a finite space. Next we show that P_r can be approximated by a sequence of discrete, stationary distributions $(P_{r;k})_{k \in \mathbb{N}}$ with $P_{r;k}$ on $\Xi_{r;k}^{\mathbb{Z}}$, where $\Xi_{r;k} = \{v_0, \dots, v_{k-1}\}$. Choose v_i as one-dimensional P_r-quantiles,

$$v_i = F_{P_r}^{-1}(\frac{i + 1/2}{k}),\ i = 0, \dots, k-1,$$

where $F_{P_r}(y) = \mathbb{P}_{P_r}[Y_t \leq y]$. We then partition the space $\mathbb{R}$ as

$$\mathbb{R} = \cup_{i=0}^{k-1} I_{v_i},$$
$$I_{v_i} = (F_{P_r}^{-1}(\frac{i}{k}), F_{P_r}^{-1}(\frac{i+1}{k})],\ i = 0, \ldots, k-2,\ I_{v_{k-1}} = (F_{P_r}^{-1}(\frac{k-1}{k}), \infty).$$

Our partition here is in terms of 'mid-values' v_i rather than some ordered numbers in $\{0, \ldots, k-1\}$ as in (2.2). Every $y \in \mathbb{R}^m$ is then an element of $I_{\iota(y_1)} \times \ldots \times I_{\iota(y_m)}$ with $y_i \in I_{\iota(y_i)}$, $\iota : \mathbb{R} \to \Xi_{r;k}$. Let

$$\nu : \mathbb{R}^{\mathbb{Z}} \to \Xi_{r;k}^{\mathbb{Z}} \subseteq \mathbb{R}^{\mathbb{Z}},\ y \mapsto (\iota(y_t))_{t \in \mathbb{Z}}.$$

Then define

$$P_{r;k} = P \circ \nu^{-1}.$$

We also extend $P_{r;k}$ from $\Xi_{r;k}^{\mathbb{Z}}$ to $\mathbb{R}^{\mathbb{Z}}$ by defining

$$P_{ext,k} = P_{r;k} \circ \nu. \tag{A.1}$$

In the sequel, we always write P_k instead of $P_{ext,k}$. Abbreviate by $Q^{(m)} = Q \circ \pi_{t_1,\ldots,t_m}^{-1}$ for any distribution Q on $\mathbb{R}^{\mathbb{Z}}$. Define '$\leq$' componentwise and let $y \in \mathbb{R}^m$. Then,

$$F_{P_{r;k}^{(m)}}(y) = \sum_{z \leq y, z \in \Xi_{r;k}^m} P_{r;k}^{(m)}(z) = F_{P_r^{(m)}}(y) + \gamma(y),$$

where

$$\gamma(y) = F_{P_{r;k}^{(m)}}(y_{\Xi_{r;k}}) - F_{P_r^{(m)}}(y)$$

with $y_{\Xi_{r;k}^m}$ being the closest point in $\Xi_{r;k}^m$ to y such that $y_{\Xi_{r;k}^m} \leq y$. But

$$|\gamma(y)| \leq \sum_{i=1}^m P_r^{(1)}(((y_{\Xi_{r;k}^m})_i, y_i]) \to 0\ (k \to \infty),$$

since $P_r^{(1)}$ is continuous and $|(y_{\Xi_{r;k}^m})_i - y_i| \to 0\ (k \to \infty)$. Thus, we have shown

$$P_{r;k} \circ \pi_{t_1,\ldots,t_m}^{-1} \Rightarrow P_r \circ \pi_{t_1,\ldots,t_m}^{-1}\ (k \to \infty)\ \forall t_1, \ldots, t_m \in \mathbb{Z},\ \forall m \in \mathbb{N} \tag{A.2}$$

with $P_{r;k}$ stationary on $\Xi_{r;k}^{\mathbb{Z}}$. Finally, set $P_k = P_{s;k} + P_{r;k}$. Then, by the first argument in Step 1 and (A.2) we get

$$P_k \circ \pi_{t_1,\ldots,t_m}^{-1} \Rightarrow P \circ \pi_{t_1,\ldots,t_m}^{-1}\ (k \to \infty)\ \forall t_1, \ldots, t_m \in \mathbb{Z},\ \forall m \in \mathbb{N} \tag{A.3}$$

with P_k stationary on $\Xi_k^{\mathbb{Z}}$, $\Xi_k = \Xi_{s;k} \cup \Xi_{r;k}$.

Step 2. Show that P_k on $\Xi_k^{\mathbb{Z}}$ can be approximated by a sequence of stationary, ergodic Markov chains $(P_{k,\ell})_{\ell \in \mathbb{N}}$ on $\Xi_k^{\mathbb{Z}}$. $P_{k,\ell}$ will be constructed as a Markov chain of order

$p = p_{k,\ell} = p_k \to \infty$ $(k \to \infty)$. We denote for a set E in Ξ_k and $z_{-p+1}^0 \in \Xi_k^p$ the transition kernel as $\overline{P}_k(E|z_{-p+1}^0) = \sum_{z_1 \in E} \overline{P}_k(z_1|z_{-p+1}^0)$, where

$$\overline{P}_k(z_1|z_{-p+1}^0) = \begin{cases} \frac{P_k(z_{-p+1}^1)}{P_k(z_{-p+1}^0)} & \text{if } P_k(z_{-p+1}^0) \neq 0 \\ 0 & \text{if } P_k(z_{-p+1}^0) = 0 \end{cases}$$

Now, modify $\overline{P}_k(.|.)$, denoted by $P_{k,\ell}(.|.)$ such that

$$\max_{E, z_{-p+1}^0, \tilde{z}_{-p+1}^0} |P_{k,\ell}(E|z_{-p+1}^0) - P_{k,\ell}(E|\tilde{z}_{-p+1}^0)| < 1,$$
$$\max_{z_{-p+1}^1} |P_{k,\ell}(z_1|z_{-p+1}^0) - \overline{P}_k(z_1|z_{-p+1}^0)| \leq \ell^{-1}, \tag{A.4}$$

Then, $P_{k,\ell}(.|.)$ generates a stationary ergodic Markov chain of order $p = p_k$ on $\Xi_k^{\mathbb{Z}}$ which we denote by $P_{k,\ell}$, cf. Doukhan (1994). Moreover, the stationary distribution of any p-tuple z_1^p is given by

$$P_{k,\ell}(z_1^p) = g_{z_1^p}(P_{k,\ell}(.|.)), \tag{A.5}$$

where all the g's are continuous (in fact, they are solutions of a linear system of equations). On the other hand for the stationary P_k we have,

$$P_k(z_1^p) = \sum_{z_0} \overline{P}_k(z_p|z_1^{p-1}z_0)P_k(z_1^{p-1}z_0)$$

and hence by (A.4),

$$P_k(z_1^p) = \sum_{z_0} P_{k,\ell}(z_p|z_1^{p-1}z_0)P_k(z_1^{p-1}z_0) + \gamma_{k,\ell},\ \gamma_{k,\ell} \to 0\ (\ell \to \infty).$$

This is approximately of the form defining the stationary probabilities in the Markov chain $P_{k,\ell}$, implying that

$$P_k(z_1^p) = g_{z_1^p}(P_{k,\ell}(.|.)) + \tilde{\gamma}_{k,\ell},\ \tilde{\gamma}_{k,\ell} \to 0\ (\ell \to \infty)$$

with g's as in (A.5). Then, by (A.5),

$$P_{k,\ell}(z_1^p) \to P_k(z_1^p)\ (\ell \to \infty)\ \forall z_1^p \in \Xi_k^p.$$

Thus, by extending $P_{k,\ell}$ and P_k to $\mathbb{R}^{\mathbb{Z}}$ as in (A.1) in Step 1,

$$P_{k,\ell} \circ \pi_{t+1,\ldots,t+m}^{-1} \Rightarrow P_k \circ \pi_{t+1,\ldots,t+m}^{-1}\ (\ell \to \infty)\ \forall t \in \mathbb{Z},\ \forall 1 \leq m \leq p = p_{k,\ell} = p_k. \tag{A.6}$$

Step 3. Show that $P_{k,\ell}$ on $\Xi_k^{\mathbb{Z}}$ can be approximated by a sequence $(P_{k,\ell,m})_{m \in \mathbb{N}}$ of stationary, ergodic QVLMC's with $P_{k,\ell,m}$ on $\mathbb{R}^{\mathbb{Z}}$. The process $(Z_{t;k,\ell})_{t \in \mathbb{Z}} \sim P_{k,\ell}$ with $Z_{t;k,\ell}$ in Ξ_k can be approximated by

$$(Y_{t;k,\ell,m})_{t \in \mathbb{Z}} = (Z_{t;k,\ell})_{t \in \mathbb{Z}} + (\varepsilon_{t;m})_{t \in \mathbb{Z}} \sim P_{k,\ell,m},$$

with $(\varepsilon_{t;m})_{t \in \mathbb{Z}}$ a sequence of i.i.d. Uniform$([-1/(2m), 1/(2m)]$ variables, independent of $(Z_{t;k,\ell})_{t \in \mathbb{Z}}$. Then, for m sufficiently large, $P_{k,\ell,m}$ is a stationary, ergodic QVLMC of order

$p_{k,\ell,m} = p_k$ and with $q : \mathbb{R} \to \{0, \ldots, N-1\}$ ($N = |\Xi_k|$). This, because $\mathrm{supp}(\varepsilon_{t;m}) = [-1/(2m), 1/(2m)] \to 0$ ($m \to \infty$) and by the construction in Steps 1 and 2. Moreover, finite-dimensional weak convergence of $P_{k,\ell,m}$ to $P_{k,\ell}$ follows immediately by the definition of $\varepsilon_{t;m}$. This, together with (A.3) and (A.6) completes the proof by using a diagonal argument to choose a sequence $(P_n)_{n\in\mathbb{N}}$ from $(P_{k,\ell,m})_{k,\ell,m\in\mathbb{N}}$. □

Proof of Theorem 2.2. Observe that for a QSVLMC with context function $\tilde{c}(.)$,

$$\mathbb{P}[q(Y_1) = x_1|x_{-\infty}^0] = \int_{I_{x_1}} f(y_1|\tilde{c}(x_{-\infty}^0))dy_1 = \mathbb{P}[q(Y_1) = x_1|\tilde{c}(x_{-\infty}^0)] \ \forall x_{-\infty}^1 \in \mathcal{X}^\infty.$$

Therefore, $(X_t)_{t\in\mathbb{Z}}$ with $X_t = q(Y_t)$ is a VLMC with context function $c(.)$ such that $|c(x_{-\infty}^0)| \leq |\tilde{c}(x_{-\infty}^0)|$ for all $x_{-\infty}^0$. This proves assertion (i). (Note that it is possible to construct a QSVLMC process with $|c(x_{-\infty}^0)| < |\tilde{c}(x_{-\infty}^0)|$ for some $x_{-\infty}^0$).

For assertion (ii) we have by assumption (B),

$$f(y_1|x_{-\infty}^0) = \sum_{x_1\in\mathcal{X}} f(y_1|x_1)P_c(x_1|c(x_{-\infty}^0))$$

with $c(.)$ the context function of the quantized VLMC $(X_t)_{t\in\mathbb{Z}}$, as described by (i). Clearly, $c(.) = \tilde{c}(.)$ with $\tilde{c}(.)$ the context function of the QSVLMC. Moreover, the formula above characterizes a QVLMC: thus $(Y_t)_{t\in\mathbb{Z}}$ is a QVLMC with context function $\tilde{c}(.)$ equal to the context function $c(.)$ of the quantized VLMC $(X_t)_{t\in\mathbb{Z}}$. □

Proof of Proposition 3.1. Observe that by assumption (C) and assuming $\hat{f}_{X_t}(Y_t) \neq 0$ for $t = p+1, \ldots, n$ (i.e., the non-trivial case),

$$\log(\hat{P}_{\tau_{c_i}}(Y_1^n)) = \sum_{t=p+1}^{n} \log(\hat{f}_{X_t}(Y_t)) + \sum_{t=p+1}^{n} \log(\hat{P}(X_t|c_i(X_{t-p}^{t-1}))) \ (i = 1, 2).$$

Therefore,

$$\tilde{\Delta}_{\tau_{c_1},\tau_{c_2}}(Y_1^n) = \sum_{t=p+1}^{n} \left(\log(\hat{P}(X_t|c_1(X_{t-p}^{t-1}))) - \log(\hat{P}(X_t|c_2(X_{t-p}^{t-1})))\right).$$

□

Justification of formula (4.1). Since the quantized process $(X_t)_{t\in\mathbb{Z}}$ is a VLMC, the result follows immediately from Bühlmann & Wyner (1997) (here, the true underlying VLMC is fixed and the conditions (A1)-(A3) from Bühlmann & Wyner (1997) are met by our assumption (D1)).

Proof of formula (4.2). By (4.1) it is easy to see that

$$n^{1/2}\{\hat{P}(x_1|\hat{c}(x_{-\infty}^0)) - P_c(x_1|c(x_{-\infty}^0))\} = n^{1/2}\{\hat{P}(x_1|c(x_{-\infty}^0)) - P_c(x_1|c(x_{-\infty}^0))\} + o_P(1).$$

Since the VLMC $(X_t)_{t\in\mathbb{Z}}$ is geometrically ϕ-mixing (see Proposition 2.1, the result follows then from well known results for the maximum likelihood estimator $\hat{P}(x_1|c(x_{-\infty}^0))$ in a finite state Markov chain, cf. Basawa & Rao (1980, Ch.2.2). For $x_1, x_2 \in \mathcal{X}$, $w_1, w_2 \in \tau_c$, the inverse Fisher-information is given by

$$I^{-1}(x_1w_1, x_2w_2) = \delta_{w_1,w_2}\frac{1}{P(w_1)}(\delta_{x_1,x_2}P(x_1|w_1) - P(x_1|w_1)P(x_2|w_2)). \tag{A.7}$$

Proof of formula (4.3). By a Taylor expansion we get

$$
\begin{aligned}
& n^{1/2}(\hat{\mathbb{P}}[Y_t \in E|X_t = x] - \mathbb{P}[Y_t \in E|X_t = x]) \\
= \quad & P(x)^{-1}n^{-1/2}\sum_{t=1}^{n}(1_{[Y_t\in E, X_t=x]} - \mathbb{P}[Y_t \in E, X_t = x]) \\
- \quad & \mathbb{P}[Y_t \in E|X_t = x]P(x)^{-1}n^{1/2}(n^{-1}N(x) - P(x)) + o_P(1).
\end{aligned}
$$

Now asymptotic normality follows by the geometric ϕ-mixing property of $(Y_t)_{t\in\mathbb{Z}}$, see Proposition 2.1. The asymptotic variance $\nu_{as}^2(E,x)$ is given by

$$
\begin{aligned}
& \nu_{as}^2(E,x) = \nu_1^2 + \nu_2^2 - 2\nu_{1,2}, \\
& \nu_1^2 = P(x)^{-2}\sum_{k=-\infty}^{\infty}\mathrm{Cov}(1_{[Y_0\in E, X_0=x]}, 1_{[Y_k\in E, X_k=x]}), \\
& \nu_2^2 = P(x)^{-2}\mathbb{P}[Y_t \in E|X_t = x]^2\sum_{k=-\infty}^{\infty}\mathrm{Cov}(1_{[X_0=x]}, 1_{[X_k=x]}), \\
& \nu_{1,2} = P(x)^{-2}\mathbb{P}[Y_t \in E|X_t = x]\sum_{k=-\infty}^{\infty}\mathrm{Cov}(1_{[Y_0\in E, X_0=x]}, 1_{[X_k=x]}).
\end{aligned} \tag{A.8}
$$

Proof of formula (4.4). Since $\hat{\mathbb{P}}[Y_t \in E] = n^{-1}\sum_{t=1}^{n}1_{[Y_t\in E]}$ we obtain the result by the geometric ϕ-mixing property for $(Y_t)_{t\in\mathbb{Z}}$, see Proposition 2.1. The asymptotic variance is given by

$$\xi_{as}^2(E) = \sum_{k=-\infty}^{\infty}\mathrm{Cov}(1_{[Y_0\in E]}, 1_{[Y_k\in E]}). \tag{A.9}$$

Proof of formula (4.5). By the geometric ϕ-mixing property of $(Y_t)_{t\in\mathbb{Z}}$, see Proposition 2.1, we obtain

$$n^{1/2}(\overline{Y}_x - \mathbb{E}[Y_t|X_t = x]) \Rightarrow \mathcal{N}(0, \zeta_1^2(x)) \ (n \to \infty), \tag{A.10}$$

where $\zeta_1^2(x)$ has an analogous formula to (A.8), replacing $1_{[Y_t\in E, X_t=x]}$ by $Y_t 1_{[X_t=x]}$ and $\mathbb{P}$ by $\mathbb{E}$. By formulae (3.8), (A.10) and (4.2) we obtain the required result. The value $\zeta_{as}^2(m,s)$ is a linear combination of covariances in (A.7) and of the expression in (A.10).

Justification of formula (4.6). Assumption (D1), implying the geometric ϕ-mixing property for $(Y_t)_{t\in\mathbb{Z}}$, says that $n^{-1}N(x) - P(x) = O_P(n^{-1/2})$ converges faster than the (unconditional) kernel density estimator. We can thus asymptotically replace $n^{-1}N(x)$ by $P(x)$. Moreover, since $K(.)$ has compact support and $y \in \overset{\circ}{I}_x$ we can asymptotically replace $(nh^d)^{-1}\sum_{t=1}^{n}K(\frac{y-Y_t}{h})1_{[X_t=x]}$ by $\tilde{f}(y) = (nh^d)^{-1}\sum_{t=1}^{n}K(\frac{y-Y_t}{h})$. We thus asymptotically replace $\hat{f}_x(y)$ by $\frac{1}{P(x)}\tilde{f}(y)$. Note also that for $y \in \overset{\circ}{I}_x$, $f_x(y) = f(y)/P(x)$ with $f(y) = \frac{d}{dt}\mathbb{P}[Y_t \le y]$. Regularity conditions for $f_x(.)$ then carry over to $f(.)$. Assuming the regularity assumptions preceding (4.6), the results in Boente & Fraiman (1995) imply

$$(nh^d)^{1/2}(\tilde{f}(y) - f(y)) \Rightarrow \mathcal{N}(B_{as}(y), \sigma_{as}^2(y)),$$

$$B_{as}(y) = \lim_{n\to\infty} n^{1/2}h^{2+d/2}\int_{\mathbb{R}^d} K(z)z'Hf(y)zdz/2,\ Hf(y) = \frac{\partial^2 f}{\partial y_i \partial y_j}(y),$$
$$\sigma^2_{as}(y) = \int_{\mathbb{R}^d} K^2(z)dzf(y).$$

By the asymptotic replacement of $\hat{f}_x(y)$ with $\frac{1}{P(x)}\tilde{f}(y)$ we get the convergence in (4.6) with

$$B_{as}(x,y) = \lim_{n\to\infty} n^{1/2}h^{2+d/2}\int_{\mathbb{R}^d} K(z)z'Hf_x(y)zdz/2,\ Hf_x(y) = \frac{\partial^2 f_x}{\partial y_i \partial y_j}(y),$$
$$\sigma^2_{as}(x,y) = \int_{\mathbb{R}^d} K^2(z)dz\frac{f_x(y)}{P(x)}. \tag{A.11}$$

Acknowledgments: I thank H. Künsch for interesting comments.

References

[1] Basawa, I.V. & Rao, B.L.S.Prakasa (1980). Statistical Inference for Stochastic Processes. Academic Press, New York.

[2] Boente, G. & Fraiman, R. (1995). Asymptotic distribution of data-driven smoothers in density and regression estimation under dependence. Canadian Journal of Statistics **23** 383-397.

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

[4] Bühlmann, P. (1997). Model selection for variable length Markov chains and tuning the context algorithm. Research Report 82, Seminar für Statistik, ETH Zürich.

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

[6] Doukhan, P. (1994). Mixing. Properties and Examples. Lecture Notes in Statistics **85**. Springer.

[7] Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics **7** 1-26.

[8] Friedman, J.H. (1991). Multivariate adaptive regression splines. Annals of Statistics **19** 1-50.

[9] Gersho, A. & Gray, R.M. (1992). Vector Quantization and Signal Compression. Kluwer.

[10] Karpoff, J.M. (1987). The relation between price changes and trading volume: a survey. J. of Financial and Quantitative Analysis **22** 109-126.

[11] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. Annals of Statistics **17** 1217-1241.

[12] Lewis, P.A.W. & Stevens, J.G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). J. of the American Statistical Association **86** 864-877.

[13] McCullagh, P. & Nelder, J.A. (1989). Generalized Linear Models. Second edition. Chapman and Hall.

[14] Nobel, A.B. (1997). Recursive partitioning to reduce distortion. IEEE Transactions on Information Theory **IT-43** 1122-1133.

[15] Priestley, M.B. (1988). Non-linear and Non-stationary Time Series Analysis. Academic Press.

[16] Rissanen, J.J. (1983). A universal data compression system. IEEE Transactions on Information Theory **IT-29** 656-664.

[17] Shephard, N.G. (1996). Statistical aspects of ARCH and stochastic volatility. In Time Series Models in Econometrics, Finance and other Fields (Eds. D.R. Cox, D.V. Hinkley & O.E. Barndorff-Nielsen), pp. 1-67. Chapman and Hall.

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

[19] Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall.

[20] Tjøstheim, D. (1994). Non-linear time series: a selective review. Scandinavian J. of Statistics **21** 97-130.

[21] Tong, H. (1990). Non-linear Time Series. A Dynamical System Approach. Oxford University Press.

[22] Weinberger, M.J., Rissanen, J.J. & Feder, M. (1995). A universal finite memory source. IEEE Transactions on Information Theory **IT-41** 643-652.

Seminar für Statistik
ETH Zentrum, HG G32.2
CH-8092 Zürich
Switzerland
E-mail: buhlmann@stat.math.ethz.ch