Dynamic Adaptive Partitioning for Nonlinear Time Series

Peter Bühlmann

Seminar für Statistik, ETH Zürich, CH-8092 Zürich, Switzerland
buhlmann@stat.math.ethz.ch

September 1998

Summary

We propose a dynamic adaptive partitioning scheme for nonparametric analysis of stationary nonlinear time series. It yields estimates of the whole probability distribution of the underlying process. 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 new. We make it constructive by an approach based on quantisation of the data and adaptively modelling partition cells with a parsimonious Markov chain. The methodology is formulated in terms of a new model class, the so-called quantised variable length Markov chains. It is a new extension of finite-valued variable length Markov chains to processes with values in $\mathbb{R}^d$. We discuss estimation, explore asymptotic properties of the new method and give some numerical results which reflect the finite sample behaviour.

Some key words: Conditional heteroscedasticity; Context algorithm; Markov chain; Multivariate time series; ϕ-mixing; Prediction; Quantisation; Stationary process; Tree model.

Short title: Dynamic adaptive partitioning

1 Introduction

Nonparametric methods which are able to adapt to local sparseness of the data are often substantially better than non-adaptive procedures because of the curse of dimensionality, and estimation of the mean as a function of predictor variables with adaptive partitioning schemes has attracted much attention (Breiman et al., 1984; Friedman, 1991; Gersho & Gray, 1992). Some of these schemes have been studied also in the case of stationary time series (Lewis & Stevens, 1991; Nobel, 1997), but none of the schemes use the simple fact, that in the case of a time series, the partition cells themselves typically have a dynamic characteristic. Consider a stationary real-valued pth-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 uses models of the form

$$E_{part}(Y_t|S_{t-1}) = \sum_{j=1}^{J} c_j I(S_{t-1} \in R_j) \tag{1.1}$$

where $\{R_j; j = 1, \ldots, J\}$ is a partition of the state space $\mathbb{R}^p$. This is the common model in the regression set-up with independent errors. The various schemes differ by adaptively producing different partitions. However, for the time series case we make use of the following facts:

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

(2) the partition cells containing the state vector $V_{t-1} = \sum_{j=1}^{J} R_j I(S_{t-1} \in R_j)$ ($t \in \mathbb{Z}$) form 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.

Facts (1) and (2) say that we can learn partially about a future Y_t via the future partition cell V_t from the partition cell process $V_1, \ldots, V_{t-1}$ and the data $Y_1, \ldots, Y_{t-1}$. This explains the expression 'dynamic adaptive partitioning' in the title. The novel approach here is additionally to model the partition cell process $\{V_t\}_t$, thus 'making dependence our friend' for adaptive partitioning. We propose quantisation and parsimonious Markov modelling. Both can be described in terms of a new model class for stationary, ergodic time series with values in $\mathbb{R}^d$ ($d \geq 1$), the so-called quantised variable length Markov chains, which are new extensions of Markov chains with variable length memory from finite spaces to $\mathbb{R}^d$-valued variables. The finite space case is known as 'tree model', 'FSMX model', 'finite-memory source' or 'variable length Markov chain'; see Rissanen (1983), Weinberger, Rissanen & Feder (1995) and Bühlmann & Wyner (1997). A main focus of the paper is on estimation of the whole distribution of stochastic processes in the new class of Markov chains, a much more general task than nonparametric estimation of conditional expectations with CART or MARS for example.

Of particular importance is a robustness property of the new chains against model misspecification. We argue in §2.4 that every stationary process can be approximated by a quantised variable length Markov chain and in §§3.2 and 3.5, that we are able to find and fit an appropriate member of this class of chains. We provide a financial illustration. Also of great interest in risk management of financial assets are measures such as conditional variances, i.e. volatility, the conditional quantiles or the conditional expected shortfall $E(Y_t|Y_t \le c_{t-1}, Y_{t-1}, Y_{t-2}, \ldots)$ with $c_{t-1} \in \mathbb{R}$ a quantile given the past up to time $t-1$; see McNeil & Frey (1998). The general aim is the knowledge of the conditional distribution given the past. Our approach yields consistent estimators thereof, essentially requiring only stationarity of the data. We analyse in §4.2 some risk questions for daily returns of the BMW stock price.

2 The Quantised variable length Markov chain model

2.1 Introduction

Our general strategy for fitting a nonlinear time series model is to quantise the data and then use an adaptively estimated parsimonious Markov model for the quantised series. The issues of choosing both the amount of quantisation and a good model are addressed in §3.5.

We assume the data $Y_1, \ldots, Y_n$ are 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 quantiser of $\mathbb{R}^d$ into a categorical set $\mathcal{X} = \{0, 1, \ldots, N-1\}$, inducing a partition,

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

2.2 Variable length Markov chains for categorical variables

Consider a stationary process $\{X_t\}_t$ with values in a finite categorical space $\mathcal{X} = \{0, 1, \ldots, N-1\}$ as in (2.1). 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 vector whose components are written in reverse order. First, we define variable length Markov chains, which are related to tree models, FSMX models and finite-memory sources; see section §1 for references.

DEFINITION 1 *Let $\{X_t\}_t$ 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*

$$\ell = \min\{k; pr(X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0) = pr(X_1 = x_1 | X_{-k+1}^0 = x_{-k+1}^0) \textit{ for all } x_1 \in \mathcal{X}\},$$

where $\ell \equiv 0$ corresponds to independence. Then, $c(.)$ is called a context function and, for any $t \in \mathbb{Z}$, $c(x_{-\infty}^{t-1})$ is called the context for the variable x_t.

The name context refers to the portion of the past that determines the probability for 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 *Let $\{X_t\}_t$ be a stationary process with values $X_t \in \mathcal{X}$ and corresponding context function $c(.)$ as given in Definition 1. Let p be the smallest integer such that*

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

Then, $c(.)$ is called a context function of order p, and if $p < \infty$, $\{X_t\}_t$ is called a variable length Markov chain of order p.

We sometimes identify a variable length Markov chain $\{X_t\}_t$ with its probability distribution P_c on $\mathcal{X}^{\mathbb{Z}}$. Also, we often write $P_c(x_i^j) = \mathrm{pr}(X_i^j = x_i^j)$ and $P_c(x_j|x_i^{j-1}) = \mathrm{pr}(X_j = x_j|X_i^{j-1} = x_i^{j-1})$ $(i < j)$ for $\{X_t\}_t \sim P_c$.

Clearly, a variable length Markov chain of order p is a Markov chain of order p, now having a memory of variable length ℓ. By requiring stationarity, a variable length Markov chain 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 class of variable length Markov chains contains many more models than the class of full Markov chains, and it is in this sense richer. It generally allows a finer trade-off between bias and variance, and typically yields a better strategy for dealing with the curse of dimensionality.

A variable length Markov chain is a tree-structured model with a root node on top, from which the branches grow downwards, so that every internal node has at most $N = |\mathcal{X}|$ offspring. 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 $x_{-\ell(x_{-\infty}^0)+1}$. Note that such context trees do not have to be complete, i.e. every internal node does not need to have exactly $N = |\mathcal{X}|$ offspring.

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

Figure 1 about here.

DEFINITION 3 *Let $c(.)$ be a context function of a variable length Markov chain of order p. The context tree τ and terminal node context tree τ^T are defined as*

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

Definition 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 variable length Markov chain with context function $c(.)$. An internal node with $b < N = |\mathcal{X}|$ offspring implicitly adds one complementary offspring, lumping the $N-b$ absent offspring together to a single new terminal node w_{new} which represents a single state in τ_c.

2.3 Quantised variable length Markov chains for $\mathbb{R}^d$-valued variables

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

$$\{X_t\}_t \text{ is an } \mathcal{X}\text{-valued variable length Markov chain.} \tag{2.3}$$

Given $X_t = x$, define Y_t independently of $\{Y_s, X_s;\ s \neq t\}$, by

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

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

DEFINITION 4 *The process $\{Y_t\}_t$ defined by (2.3) and (2.4) is called a quantised variable length Markov chain.*

By formula (2.4), a quantised variable length Markov chain has the property that its quantised values, with the correct quantiser q, form a variable length Markov chain, i.e. $\{q(Y_t)\}_t = \{X_t\}_t$ is a variable length Markov chain. Also, a quantised variable length Markov chain $\{Y_t\}_t$ is a stationary $\mathbb{R}^d$-valued Markov chain, generally of higher order, with a memory induced by the underlying variable length Markov chain,

$$\begin{aligned} \mathrm{pr}(Y_t \leq y | Y_{-\infty}^{t-1}) &= \sum_{x\in\mathcal{X}} \int_{-\infty}^{y} f_x(z)dz\ \mathrm{pr}\{X_t = x | c(X_{-\infty}^{t-1})\} \\ &= \mathrm{pr}\{Y_t \leq y | c(X_{-\infty}^{t-1})\},\ y \in \mathbb{R}^d,\ X_s = q(Y_s), \end{aligned}$$

where '$\leq$' is defined componentwise. However, the minimal state space of $\{Y_t\}_t$ has specific structure and is the same as for $\{X_t\}_t$, namely τ_c as given in Definition 3.

For the univariate quantised variable length Markov chain model, the quantiser $q : \mathbb{R} \to \mathcal{X} = \{0, 1, \ldots, N-1\}$ in (2.1) and (2.2) is usually chosen in terms of disjoint intervals in $\mathbb{R}$; see formula (3.1). For the multivariate model, the quantiser is $q : \mathbb{R}^d \to \mathcal{X} = \{0, 1, \ldots, N-1\}$. General vector quantisation is less interpretable than scalar quantisation, particularly in terms of individual series. We propose, but do not require, scalar quantisation of different individual time series,

$$\begin{aligned} &q : \mathbb{R}^d \to \mathcal{X},\ q(Y_t) = \{q_1(Y_{1,t}), \ldots, q_d(Y_{d,t})\},\ Y_t = (Y_{1,t}, \ldots, Y_{d,t}), \\ &q_j : \mathbb{R} \to \mathcal{X}_j = \{0, 1, \ldots, N_j\}\ (j = 1, \ldots, d),\ \mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d, \end{aligned} \tag{2.5}$$

with a product space $\mathcal{X}$, labelled arbitrarily by $0, 1, \ldots, N-1$ with $N = N_1 \cdots N_d$. The flexibility of quantised variable length Markov chains also allows us to model multivariate time series data with some real-valued and some categorical components.

2.4 Properties of quantised variable length Markov chains

The dynamic property of a quantised variable length Markov chain is given by the variable length Markov chain model of the quantised series. Since the variables Y_t given $\{X_s\}_s$ are independent and depend only on their quantised values X_t, stationarity and ergodicity of $\{Y_t\}_t$ is inherited from $\{X_t\}_t$. Note that this statement is meant to be unconditional on $\{X_t\}_t$. A sufficient condition for ergodicity is then implied by a Doeblin-type condition, and stationarity is already implicitly assumed by our Definitions 2 and 4.

Assumption A. The underlying variable length Markov chain $\{X_t\}_t \sim P_c$ on $\mathcal{X}^{\mathbb{Z}}$ satisfies

$$\min_{x \in \mathcal{X}, w \in \tau_c} P_c(x|w) > 0.$$

PROPOSITION 1 *Let $\{Y_t\}_t$ be a quantised variable length Markov chain as given in Definition 4, satisfying assumption A. Then, $\{Y_t\}_t$ is ergodic and uniformly mixing with mixing coefficients satisfying $\phi(i) \leq const.\rho^i$ for all $i \in \mathbb{N}$, where $0 < \rho < 1$.*

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

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. However, this does not mean that only very short range phenomena could be modelled with quantised variable length Markov chains. Indeed, Theorem 1 below discusses the breadth of the model class, which is weakly dense in the set of stationary, $\mathbb{R}^d$-valued processes. 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 consequently by $P \circ \pi^{-1}_{t_1,\ldots,t_m}$ the m-dimensional distribution of $(Y_{t_1}, \ldots, Y_{t_m})$, where $\{Y_t\}_t$ has distribution P on $(\mathbb{R}^d)^{\mathbb{Z}}$. Moreover, we denote by '$\Rightarrow$' weak convergence.

THEOREM 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 ergodic, $\mathbb{R}^d$-valued quantised variable length Markov chains, such that*

$$P_n \circ \pi^{-1}_{t_1,\ldots,t_m} \Rightarrow P \circ \pi^{-1}_{t_1,\ldots,t_m} \;\; \textit{as } n \to \infty, \;\; \textit{for all } t_1, \ldots, t_m \in \mathbb{Z}, \;\; \textit{for all } m \in \mathbb{N}.$$

A sketch of the proof is given in the Appendix. For smooth P, a coarse quantisation in the quantised variable length Markov chain is expected to work well.

2.5 Prediction with quantised variable length Markov chains

Since a quantised variable length Markov chain specifies the whole probability distribution of the process, any predictor can be computed in such a model. The general formula for the m-step-ahead conditional density of Y_{n+m} given $Y^n_{-\infty}$ is

$$f_{Y_{n+m}|Y^n_{-\infty}}(y) = \sum_{x^{n+m}_{n+1} \in \mathcal{X}^m} f_{x_{n+m}}(y) \prod_{j=0}^{m-1} P_c\{x_{n+m-j} | c(x^{n+m-j-1}_{n+1} X^n_{-\infty})\}, \tag{2.6}$$

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 quantised variable length Markov chain thus models the conditional density as a function of finitely many of the past quantised values $X_{-\infty}^{n}$ rather than $Y_{-\infty}^{n}$. As with any partitioning scheme, the predictor in (2.6) then ranges over only a finite, although often large, number of different densities. When we specialise to the optimal mean square error m-step ahead predictor in a quantised variable length Markov chain, it is easy to see that, for a fixed function $g : \mathbb{R}^d \to \mathbb{R}^q$ $(d, q \in \mathbb{N})$,

$$\begin{aligned} &E\{g(Y_{n+m})|Y_{-\infty}^{n}\} = E\{g(Y_{n+m})|c(X_{-\infty}^{n})\} \\ = &\sum_{x_{n+1}^{n+m} \in \mathcal{X}^m} E\{g(Y_{n+m})|X_{n+m} = x_{n+m}\} \prod_{j=0}^{m-1} P_c\{x_{n+m-j}|c(x_{n+1}^{n+m-j-1}X_{-\infty}^{n})\}. \end{aligned} \tag{2.7}$$

Again, this predictor only takes values in a finite, although often large, subset of $\mathbb{R}^q$. Clearly, by choosing appropriate functions $g(.)$, we obtain predictions of conditional moments given the past. An example is the conditional variance, i.e. volatility, in financial time series. With $g(y) = y$, formula (2.7) also describes the differences to the autoregressive threshold models SETAR (Tong, 1990), in which the thresholds are determined by one lagged variable, corresponding to a partition of the real line $\mathbb{R}$, and autoregressions are used within the partition. With a quantised variable length Markov chain, the partitions are given through all p lagged variables, where p is the order of the variable length Markov chain, and constants, which are mixtures of conditional means, are used within a partition.

2.6 Interpretation as dynamic adaptive partitioning

We discuss now in more detail the issues (1) and (2) from §1, for notational simplicity only for $\mathbb{R}$-valued processes. The coefficients c_j in (1.1) are constants, depending only on the index j of the partition element R_j,

$$c_j = E(Y_t|S_{t-1} \in R_j) = E(Y_t|V_{t-1} = R_j) = m_p(R_j), \; m_p : \mathcal{B}^p \to \mathbb{R}, \tag{2.8}$$

with $\mathcal{B}^p$ the Borel σ-algebra of $\mathbb{R}^p$. The fact that $\{V_t\}_t$ is a stochastic process, where information from the past could be non-trivial, is not used with such general static partitioning. It is not difficult to show that quantised variable length Markov chains induce a dynamic partitioning: a partition model holds as in (1.1) with certain partition cells $R_1, \ldots, R_J$, but the coefficients c_j are now of the form

$$\begin{aligned} c_j &= \sum_{v_t \in R_j} m_1(v_t)\mathrm{pr}_{VLMC}(V_t = v_t|V_{t-1} = R_j), \\ m_1(v_t) &= m_1(v_{1,t}) = E(Y_t|Y_t \in v_{1,t}), \; v_t = v_{1,t} \times \ldots \times v_{p,t}, \; m_1 : \mathcal{B}^1 \to \mathbb{R}, \end{aligned} \tag{2.9}$$

with $\mathcal{B}^1$ the Borel σ-algebra of $\mathbb{R}^1$ and $\mathrm{pr}_{VLMC}(V_t = v_t|V_{t-1} = v_{t-1})$ the probability induced by the variable length Markov chain $\{X_t\}_t$; i.e. $\mathrm{pr}_{VLMC}(V_t = v_t|V_{t-1} = v_{t-1}) = \mathrm{pr}(X_t = x_t|c(X_{t-p}^{t-1}) = w_j)$, 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. As described by (2.9), our dynamic partitioning models $\{V_t\}_t$ as a Markov chain and uses a function $m_1(.)$ with domain $\mathcal{B}^1$, involving only a one-dimensional structure. This is in contrast to static partitioning as described in (2.8), where no dynamic model for $\{V_t\}_t$ is assumed and a function $m_p(.)$ with p-dimensional domain $\mathcal{B}^p$ is used.

The dynamic structure of a quantised variable length Markov chain can be interpreted in terms of a context tree, see Definition 3, and a one-dimensional simple nonparametric structure for $E(Y_t|X_t)$. Clearly, a dynamic adaptive partitioning scheme can be constructed in many different ways. For instance, one might use a static scheme like CART for the partition cells $R_1, \ldots, R_J$, a full Markov chain of order 1 for the partition cell process $\{V_t\}_t$ and as with quantised variable length Markov chains, a simple nonparametric structure for the distribution of Y_t given $Y_t \in R_j$. This has the potential to be consistent for the conditional expectation $E(Y_t|Y_{-\infty}^{t-1})$ but generally not for the whole distribution of the underlying process.

3 The fitting of quantised variable length Markov chains

3.1 Choice of quantiser

We first have to find an appropriate quantiser 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:

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

This yields an interval partition of $\mathbb{R}$ with equal number of observations per partition cell. Specification of N is discussed in §3.5. In the multivariate case, we could use a quantiser as in (2.5), with q_j estimated as in (3.1) in terms of the quantiles of the jth individual series. The choice of an appropriate q or an appropriate size N of $\mathcal{X}$ could also be given by the application.

3.2 Context algorithm

Given data $X_1, \ldots, X_n$ from a variable length Markov chain P_c on $\mathcal{X}^{\mathbb{Z}}$, and if we assume that q is the correct quantiser, the aim is to find the underlying context function $c(.)$ and an estimate of P_c. In the sequel we make the convention that quantities involving time indices $t \notin \{1, \ldots, n\}$ equal zero or are irrelevant. Let

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

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

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

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

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

where $K = K_n \sim C \log(n)$ and $C > 2|\mathcal{X}| + 4$ is 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. That is, search for the context function $c_{max}(.)$ with terminal node context tree representation τ^T_{max}, see Definition 3, where τ^T_{max} is the biggest tree such that every element, or terminal node, in τ^T_{max} has been observed at least twice in the data. Thus τ^T_{max} is such that $w \in \tau^T_{max}$ implies $N(w) \geq 2$, and, for every τ^T, where $w \in \tau^T$ implies $N(w) \geq 2$, it holds that $\tau^T \preceq \tau^T_{max}$. Here, $\tau_1 \preceq \tau_2$ means that $w \in \tau_1$ implies $wu \in \tau_2$ 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, i.e. terminal node, of $\tau^T_{(0)}$ as follows; the order of examining is irrelevant. Let $c(.)$ be the context function corresponding to $\tau^T_{(0)}$ and let

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

where wu is an element 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 $K = K_n \sim C \log(n)$, $C > 2|\mathcal{X}| + 4$ and $\hat{P}(.|.)$ as defined in (3.3). The 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 $\tau_{(i-1)}, \tau^T_{(i-1)}$ for $i = 1, 2, \ldots$, 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 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 a hierarchical backward selection with the Δ_{wu} in Step2 essentially a log-likelihood-ratio statistic; see Bühlmann & Wyner (1997). Dependence on some values further back in the history is made weaker by considering deep nodes in the tree in a hierarchical way as less relevant.

Consistent estimation of the true context function $c(.)$, i.e. the tree structured minimal state space τ_c, is discussed in Weinberger et al. (1995) and in Bühlmann & Wyner (1997). Asymptotic normality of the transition probabilities $n^{1/2}[\hat{P}\{x_1|\hat{c}(x^0_{-\infty})\} - P(x_1|w)]$ for

$w = c(x^0_{-\infty})$ follows by the consistency of $\hat{c}(.)$ and the mixing property in Proposition 1. The context algorithm needs $O\{n\log(n)\}$ operations and is thus computationally fast.

The estimation of the minimal state space τ_c is done solely on the basis of the quantised 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.

Assumption B. Estimation of the minimal state space τ_c of the quantised variable length Markov chain, or of the underlying variable length Markov chain, is exclusively based on possibly multiple use of the log-likelihood-ratio statistic

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

where $\tau_{c_1},\ \tau_{c_2}$ are context trees 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 loglikelihood of an estimated quantised variable length Markov chain with given $c_i(.)$. For $i = 1, 2,\ c_i(.)$ is induced by τ_{c_i}, $\hat{f}\{Y_t|c_i(X_{t-p}^{t-1})\} = \hat{f}_{X_t}(Y_t)\hat{P}\{X_t|c_i(X_{t-p}^{t-1})\}$ is an estimate in the quantised variable length Markov chain for $f\{Y_t|c_i(X_{t-p}^{t-1})\}$ with $\hat{f}_x(.)$ consistent and $\hat{P}(.|.)$ as in (3.3), and p is the maximal order of $c_1(.)$ and $c_2(.)$.

Assumption B is quite natural in model selection. If we neglect the minor effect of different orders p for different τ_{c_i}'s, the context algorithm satisfies assumption B, because Δ_{wu} in Step 2 is essentially a log-likelihood-ratio statistic.

PROPOSITION 2 *Assume that $\{Y_t\}_t$ is a quantised variable length Markov chain with minimal state space τ_c. Then any estimate of τ_c satisfying assumption B is based solely on the quantised data $X_1, \ldots, X_n$.*

A proof is given in the Appendix. Proposition 2 then justifies the use of the context algorithm, which is only based on the quantised data, for estimation of the minimal state space τ_c. This is in contrast to many static partitioning schemes such as CART, where the predictors are used in a quantised form but the non-quantised response variables contribute to the estimation of a partition of the predictor space.

3.3 Estimation of densities and cumulative probabilities

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

$$\hat{f}_x(y) = \frac{n^{-1}h^{-d}\sum_{t=1}^{n} K(\frac{y-Y_t}{h})I(X_t = x)}{n^{-1}N(x)},\ y \in \mathbb{R}^d \tag{3.4}$$

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 $nh^{d+4} \to C$, as $n \to \infty$ for some constant $0 < C < \infty$; see for example Silverman (1986, Ch.4). Asymptotic normality of $(nh^d)^{1/2}\{\hat{f}_x(y) - f_x(y)\}$ for

$x \in \mathcal{X}$ and y an interior point of I_x follows by the mixing property in Proposition 1, if the quantiser q is assumed known.

If the cumulative probabilities of the observations are of more interest, one can directly use empirical distribution functions. We estimate $\mathrm{pr}(Y_t \in E)$ and $\mathrm{pr}(Y_t \in E|X_t = x)$ for some measurable set E by

$$\hat{\mathrm{pr}}(Y_t \in E) = n^{-1} \sum_{t=1}^{n} I(Y_t \in E),\ \hat{\mathrm{pr}}(Y_t \in E|X_t = x) = \frac{n^{-1}\sum_{t=1}^{n} I(Y_t \in E)I(X_t = x)}{n^{-1}N(x)}.$$

Asymptotic normality of $n^{1/2}\{\hat{\mathrm{pr}}(Y_t \in E) - \mathrm{pr}(Y_t \in E)\}$ and $n^{1/2}\{\hat{\mathrm{pr}}(Y_t \in E|X_t = x) - \mathrm{pr}(Y_t \in E|X_t = x)\}$ follows by the mixing property in Proposition 1, if q is assumed fixed.

3.4 Estimated predictors

The predictive density in (2.6) can be estimated by plugging in the density estimate from (3.4) and the estimated context function and transition probabilities for $\{X_t\}_t$ from Steps 3 and 4 in §3.2. For the predictor in (2.7), we estimate $E\{g(Y_t)|X_t = x\}$ by $\overline{g(Y)}_x = N(x)^{-1}\sum_{t=1}^{n} g(Y_t)I(X_t = x)$ and use again the plug-in technique:

$$\hat{E}\{g(Y_{n+m})|Y_1^n\} = \sum_{x_{n+1}^{n+m} \in \mathcal{X}^m} \overline{g(Y)}_{x_{n+m}} \prod_{j=0}^{m-1} \hat{P}\{x_{n+m-j}|\hat{c}(x_{n+1}^{n+m-j-1}X_1^n)\}. \tag{3.5}$$

See formula (2.7) for a proper definition of $x_{n+1}^{n+m-j-1}X_1^n$. This estimated predictor takes values in a finite, but usually large, subset of $\mathbb{R}^q$. Asymptotic normality of $n^{1/2}[\hat{E}\{g(Y_{n+m})|Y_1^n = y_1^n\} - E\{g(Y_{n+m})|s\}]$, where $s = c[\{q(y_t)\}_{t=1}^n]$ is the state of the variable length Markov chain at time n, follows by the consistency of $\hat{c}(.)$ and the mixing property in Proposition 1, if we assume m finite and q known.

3.5 Model selection

We select a quantiser q and estimate a minimal state space τ_c, or equivalently a context function $c(.)$, in a fully data-driven way. For simplicity and manageability we assume a Gaussian component quasilikelihood structure which sets us in a parametric set-up, although the original problem is of semi- or nonparametric nature.

We focus first on the univariate case. The loglikelihood function of a quantised variable length Markov chain, conditional on the first p observations, is

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

where p is the order of the underlying variable length Markov chain. Write $\mu_x = E(Y_t|X_t = x)$ and $\sigma_x^2 = \mathrm{var}(Y_t|X_t = x)$. Assume $f_x(y)$ to be the density of a $\mathcal{N}(\mu_x, \sigma_x^2)$ random variable, although then $\mathrm{supp}(f_x) = \mathbb{R}$. Consider the Gaussian component quasi-loglikelihood function

$$\ell_{quasi}(\theta; Y_1^n) = \sum_{t=p+1}^{n} \log[(2\pi\sigma_{X_t}^2)^{-1/2}\exp\{-(Y_t - \mu_{X_t})^2/(2\sigma_{X_t}^2)\}]$$
$$+ \sum_{t=p+1}^{n} \log[P_c\{X_t|c(X_{t-p}^{t-1})\}], \tag{3.6}$$

where $\theta = (\mu_0, \ldots, \mu_{N-1}, \pi)$ with $\pi_{wx} = \text{pr}\{X_t = x | c(X_{t-p}^{t-1}) = w\}$, $w \in \tau_c$, $x \in \{0, \ldots, N-2\}$. The maximum quasilikelihood estimator for $c(.)$ known,

$$\hat{\theta}_{MQLE} = \text{argmin}_\theta\{-\ell_{quasi}(\theta; Y_1^n)\} \tag{3.7}$$

provides the parameter values used for the estimated predictor in (3.5) with $g(y) = y$. For the prediction problem we thus can restrict our attention to the quasilikelihood function in (3.6) and the estimator $\hat{\theta}_{MQLE}$ in (3.7). The quasilikelihood function itself is not meant to describe the whole underlying distribution of the observations but rather the characteristics of the conditional expectation $E(Y_t|Y_1^{t-1})$, see McCullagh & Nelder (1989, Ch.9). In the parametric case, in which we find ourselves, a proper AIC-type criterion is of the form

$$-2\ell_{quasi}(\hat{\theta}; Y_1^n) + 2\text{dim}(\theta).$$

In our case, $\text{dim}(\theta) = N + |\tau_c|(N-1)$ with $N = |\mathcal{X}|$. If we replace σ_x^2 with $\hat{\sigma}_x^2 = \{N(x)-1\}^{-1}\sum_{t=1}^n (Y_t - \overline{Y}_x)^2 I(X_t = x)$, our model selection criterion then becomes

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

where $\hat{P}(.|.)$ is given in (3.3). The criterion thus employs a weighted quadratic loss for the quantisation effect, the loglikelihood for the dynamic variable length Markov chain part and a penalty. Note that the quantiser q enters implicitly. Theoretically we would search for the quantiser q and the context function $c(.) = c_q(.)$, which minimise $M^2(q,c)$. However, the search over all context functions becomes very quickly computationally infeasible. A remedy proposed in Bühlmann (1997), here applied to the dynamic variable length Markov chain part of the criterion, is to search for an optimal cut-off parameter K in the context algorithm; see Step 2 in §3.2. Our proposal is to choose the quantiser q and the cut-off parameter K, which minimise

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

where $\hat{c}_K$ is the estimated context function for the $\mathcal{X}$-valued variable length Markov chain, depending on K, and $\hat{P}(.|.)$ is as in (3.3). Note that, for given q, the search for an optimal cut-off K is only affected by the term $-2\sum_{t=p+1}^n \log[\hat{P}\{X_t|\hat{c}_K(X_{t-p}^{t-1})\}]$, thus being exactly the same as when tuning the context algorithm for a categorical valued variable length Markov chain studied in Bühlmann (1997). For the multivariate model, we choose q and K to minimise

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

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

4 Numerical examples

4.1 Simulated data

We study the predictive performance of our scheme for simulated univariate data by considering the simple problem of one-step-ahead prediction of observations as in (3.5) with $g(y) = y$. We consider various data-generating models representing different kinds of stationary processes. The sample size is denoted by n. We then evaluate an estimated one-step-ahead predictor for the next L observations: we do not re-estimate the predictor, which is always based on the first n observations. We calculate

$$PE = L^{-1} \sum_{t=n+1}^{n+L} (\hat{Y}_t - Y_t)^2,$$

with $\hat{Y}_t = \hat{E}(Y_t | Y_1^{t-1})$ the predictor estimated on the $Y_1, \ldots, Y_n$ and evaluated on Y_1^{t-1}, or equivalently on Y_{t-p}^{t-1} with p the dimension of the estimated state space. By the mixing property and letting the number L of predicted observations grow, the quantity PE approximates the final prediction error conditioned on the observed data, $E\{(\hat{Z}_t - Z_t)^2 | Y_1^n\}$, where $\{Z_t\}_t$ is an independent copy of $\{Y_t\}_t$ 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 always use here $L = 1000$. We compute the measure PE of actual predictive performance for the predictor in (3.5) with $g(y) = y$ for various quantisers q and cut-off parameter $K = \chi^2_{N-1;0.95}/2$, $N = |\mathcal{X}|$, which we have often found to be a reasonable value. In the examples, the quantisers $q = \hat{q}$ are estimated from the data as in (3.1). Varying over $\hat{q}$ then results in varying over $N = |\mathcal{X}|$. We give the model selection measure $M^2(q, K) = M^2(N)$ from §3.5, by our choice of $q = \hat{q}$ and K, as an estimate of predictive performance. We compare the quantised variable length Markov chain scheme with the predictor from an $AR(p)$ model with p chosen by the minimum AIC criterion, with projection pursuit autoregression (Friedman & Stuetzle, 1981) and with MARS (Friedman, 1991). The latter two methods are nonparametric but try to deal with the curse of dimensionality in an intelligent way. Note that only simple tasks like one-step-ahead prediction of observations can be properly done with projection pursuit or MARS, because we cannot get the conditional distribution given the past. We report below parts of a larger simulation study.

We consider first a nonparametric AR(2) model with an interaction term in the mean function:

$$Y_t = \{0.5 + 0.9 \exp(-2.354 Y_{t-1}^2)\} Y_{t-1} - \{0.8 - 1.8 \exp(-2.354 Y_{t-1}^2)\} Y_{t-2} + Z_t, \quad (4.1)$$

with $Z_t \sim \mathcal{N}(0, 0.425)$ independently for all t, and Z_t independent from $\{Y_s;\ s < t\}$. The model is specified so that $\mathrm{var}(Y_t) \approx 1$. Table 1 summarises the results for $n = 4000$ and $n = 500$.

Table 1 about here.

We abbreviate by 'QVLMC, N' the quantised variable length Markov chain predictor with quantiser as in (3.1) determined by N; 'oracle' is the predictor based on the true model; 'AR' is the minimum AIC linear autoregressive predictor; 'PPreg' and 'MARS' refer to the projection pursuit and MARS predictors, respectively, reported with the empirically best number of lagged variables. The number of terms in 'PPreg', chosen between 2 and 10, had very little effect. For 'MARS' we used the algorithm from the library MDA in S-Plus, available on the internet at 'http://lib.stat.cmu.edu/S/mda'. The model dimension indicates the complexity of the scheme; for an AR(p) predictor, the model dimension is $p+1$ because we apply a mean correction first. The performance in terms of prediction error ratio of the best selected quantised variable length Markov chain scheme according to $M^2(q,K)$ relative to the oracle are $0.425/0.474 = 0.90$ for $n = 4000$, and $0.425/0.592 = 0.72$ for $n = 500$. Projection pursuit autoregression predictably does slightly better in this case where the true model is nonparametric autoregressive with independent, identically distributed innovations. MARS is not competitive.

Another nonparametric AR(2) model is additive for the mean function in the lagged variables but with conditional heteroscedastic errors,

$$Y_t = 0.863\sin(4.636Y_{t-1}) + 0.431\cos(4.636Y_{t-2}) + (0.023 + 0.5Y_{t-1}^2)^{1/2}Z_t \quad (4.2)$$

with $Z_t \sim \mathcal{N}(0,1)$ independently for all t, and Z_t independent from $\{Y_s;\ s < t\}$. Again, the model is specified so that $\mathrm{var}(Y_t) \approx 1$. Table 2 summarizes the results for $n = 4000$, the notation is as in Table 1.

Table 2 about here.

The performance of the best selected quantised variable length Markov chain scheme relative to the oracle is $0.523/0.642 = 0.81$. Projection pursuit autoregression has about the same performance as the quantised variable length Markov chain scheme, and MARS is slightly worse.

We consider also a bivariate model,

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

with $\{Z_{1,t}\}_t$, $\{Z_{2,t}\}_t$ independent sequences, $Z_{1,t} \sim \mathcal{N}(0,1)$, $Z_{2,t} \sim \mathcal{N}(0,1)$ independently for all t and $Z_{1,t},\ Z_{2,t}$ independent of $\{Y_{1,s}, U_s;\ s < t\}$. The series $\{U_t\}_t$ is only auxiliary for the definition of $\{Y_{1,t}, Y_{2,t}\}_t$. 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 3.

Table 3 about here.

We abbreviate by 'QVLMC, ($N_1,\ N_2$)' the quantised variable length Markov chain predictor with quantiser as in (2.5) 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. Write $PE_j = 1000^{-1}\sum_{t=n+1}^{n+1000}(\hat{Y}_{j,t} - Y_{j,t})^2$ ($j = 1,2$) and by $PE_{tot} = PE_1 + PE_2$. The minimum AIC AR approximation is given by a bivariate AR(1), and the predictive performance of the best selected quantised variable length Markov chain scheme relative to the AR predictor is $1.996/1.224 = 1.63$, which is a big gain.

Other univariate models, such as autoregressive threshold, quantised variable length Markov chain and linear AR models have been studied as well. If the model is nonlinear, the quantised variable length Markov chain scheme generally outperforms the AR predictor. In the case of AR models, the loss of the quantised variable length Markov chain scheme relative to the AR predictor is small or moderate. In comparison to projection pursuit autoregression and MARS, the following may be concluded. If the model is nonparametric autoregressive with independent, identically distributed innovations, projection pursuit has a slight advantage and is the best, but MARS seems to have some difficulties with the non-additive structure as in Table 1. If the model is nonparametric autoregressive with heteroscedastic innovations, quantised variable length Markov chain and projection pursuit are similar; see Table 2. In this latter case with additive structure of the conditional mean, MARS becomes competitive. If the model is not simply specified by conditional first and second moments and independent, identically distributed innovations, there is some evidence that the quantised variable length Markov chain scheme often improves upon the projection pursuit or MARS predictor. Finally, the quantised variable length Markov chain 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.

4.2 Returns from BMW stock price

The quantised variable length Markov chain scheme yields much more general results than one-step-ahead prediction of observations, as we now illustrate. We consider daily returns $Y_t = \log(B_t) - \log(B_{t-1})$ $(t = 1, \ldots, n = 1000)$, where B_t denotes the BMW stock price at time t. Fig. 2 shows the data $Y_1, \ldots, Y_n$ and the next nine future values $Y_{n+1}, \ldots, Y_{n+9}$.

Figure 2 about here.

The fitted model is with $N = 9$ for the quantiser as in (3.1). Fig. 3 shows the one-step-ahead predicted densities $\hat{f}_{Y_t|Y_1^{t-1}}(.)$ $(t = n+1, n+5, \ldots, n+9)$ from the fitted model for six of the nine future days displayed at the bottom of Fig. 2, (g)-(i) display the following summary statistics of these predictive densities: the volatilities $\hat{\mathrm{var}}(Y_t|Y_1^{t-1})$, the skewnesses $\hat{E}[\{Y_t - \hat{E}(Y_t|Y_1^{t-1})\}^3|Y_1^{t-1}]$ and the expected shortfalls $\hat{E}(Y_t|Y_t \leq c_{t-1}, Y_1^{t-1})$ with c_{t-1} the estimated conditional 5% quantile of Y_t given Y_1^{t-1}.

Figure 3 about here.

The predictive densities for $t = n+2, n+3$ and $n+4$ were very similar to those for $t = n+1$ and $n+5$. The predictive densities are excellent exploratory forecasts for the extreme behaviours of the future returns 6 and 7: the quantised variable length Markov chain scheme is in this particular example able to predict both 'changes in regime', at future days 6 and 8, although it has to be admitted that this prediction of 'change in regime' was not always so successful in other datasets. The volatilities and conditional expected shortfall summary statistics reflect this forecasting behaviour in a more quantitative way, with the conditional skewness deviating from zero only on a small scale. We point out that the results here are, as fitted quantised variable length Markov chains, asymptotically robust against model-misspecification.

Appendix

Proof of Theorem 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 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 $F_{P_r}(y) = \mathrm{pr}_{P_r}(Y_t \le y)$. We then partition the space $\mathbb{R} = \cup_{i=0}^{k-1} I_{v_i}$ with

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

Our partition here is in terms of 'mid-values' v_i rather than some ordered numbers in $\{0, \dots, k-1\}$ 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 $P_{r;k} = P_r \circ \nu^{-1}$. 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 $Q^{(m)} = Q \circ \pi_{t_1,\dots,t_m}^{-1}$ for any distribution Q on $\mathbb{R}^{\mathbb{Z}}$. Define '$\le$' componentwise and let $y \in \mathbb{R}^m$. Then, for the cumulative distribution function,

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

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 $y_{\Xi_{r;k}^m} \le y$. But

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

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

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

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

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

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 $p = p_{k,\ell} \equiv p_k \to \infty$ $(k \to \infty)$. We define a transition kernel $\overline{P}_k(.|z_{-p+1}^0)$ for $z_{-p+1}^0 \in \Xi_k^p$,

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

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

$$\begin{aligned} &\min_{z_{-p+1}^1} P_{k,\ell}(z_1|z_{-p+1}^0) > 0, \\ &\max_{z_{-p+1}^1} |P_{k,\ell}(z_1|z_{-p+1}^0) - \overline{P}_k(z_1|z_{-p+1}^0)| \leq \ell^{-1}, \end{aligned} \tag{A.4}$$

Then, $P_{k,\ell}(.|.)$ generates a stationary ergodic Markov chain of order $p = p_k$ on $\Xi_k^{\mathbb{Z}}$ which we denote by $P_{k,\ell}$, see Doukhan (1994, Ch.2.4). 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, as solutions of a linear system of equations, continuous. 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, together with (A.5),

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

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

$$P_{k,\ell} \circ \pi_{1,\ldots,m}^{-1} \Rightarrow P_k \circ \pi_{1,\ldots,m}^{-1}\ (\ell \to \infty),\ \text{for all } 1 \leq m \leq p = p_{k,\ell} \equiv p_k. \tag{A.6}$$

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 ergodic QVLMC's with $P_{k,\ell,m}$ on $\mathbb{R}^{\mathbb{Z}}$; the approximating QVLMC's can be chosen with fixed, non-variable length memory. 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 an ergodic QVLMC of (non-variable) order $p_{k,\ell,m} \equiv p_k$ and with $q : \mathbb{R} \to \{0, \ldots, N-1\}$ where $N = |\Xi_k|$. 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 Proposition 2. We assume $\hat{f}_{X_t}(Y_t) \neq 0$ since $\hat{f}_x(.)$ is consistent. Observe that by assumption (B),

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

Therefore,

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

□

References

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

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

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

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

[5] Friedman, J.H. (1991). Multivariate adaptive regression splines (with Discussion). Ann. Statist. **19** 1-67 (Disc: 67-141).

[6] Friedman, J.H. & Stuetzle, W. (1981). Projection Pursuit Regression. J. of the American Statistical Association **76** 817-823.

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

[8] 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.

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

[10] McNeil, A.J. & Frey, R. (1998). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Preprint, Dept. of Mathematics, ETH Zürich, Switzerland.

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

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

[13] Silverman, B.W. (1986). Density Estimation for Statistics

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

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

method	sample size	model dimension	PE	$M^2(q, K)$
QVLMC, $N = 24$	4000	829	0.779	12381.3
QVLMC, $N = 20$	4000	932	0.700	12056.1
QVLMC, $N = 16$	4000	1126	0.558	11524.1
QVLMC, $N = 12$	4000	870	0.482	10953.9
QVLMC, $N = 9$	4000	489	0.474	10732.1
QVLMC, $N = 6$	4000	196	0.521	10833.6
AR	4000	10	0.842	–
PPreg, #(lags)=2	4000	–	0.433	–
MARS, #(lags)= 3	4000	–	0.782	–
oracle	–	–	0.425	–
QVLMC, $N = 9$	500	89	0.805	1677.0
QVLMC, $N = 7$	500	109	0.650	1637.0
QVLMC, $N = 6$	500	81	0.584	1563.1
QVLMC, $N = 5$	500	57	0.592	1488.0
QVLMC, $N = 4$	500	40	0.601	1500.6
QVLMC, $N = 3$	500	27	0.646	1548.0

Table 1: Performances for nonparametric AR(2) in(4.1).

[1]

[1] QVLMC: quantised variable length Markov chain; AR: minimum AIC linear autoregressive; PPreg: projection pursuit on lagged variables with 2 terms; MARS: MARS on lagged variables; oracle: true model.

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	–
PPreg, #(lags)=2	–	0.635	–
MARS, #(lags)=4	–	0.669	–
oracle	–	0.523	–

Table 2: Performances for nonparametric AR(2) in (4.2), $n = 4000$.

[2]

[2]QVLMC: quantised variable length Markov chain; AR: minimum AIC linear autoregressive; PPreg: projection pursuit on lagged variables with two terms; MARS: MARS on lagged variables; oracle: true model.

method	model dimension	(PE_1, 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 3: Performances for bivariate model, $n = 4000$.

[3]

[3]QVLMC: quantised variable length Markov chain; AR: minimum AIC linear autoregressive.

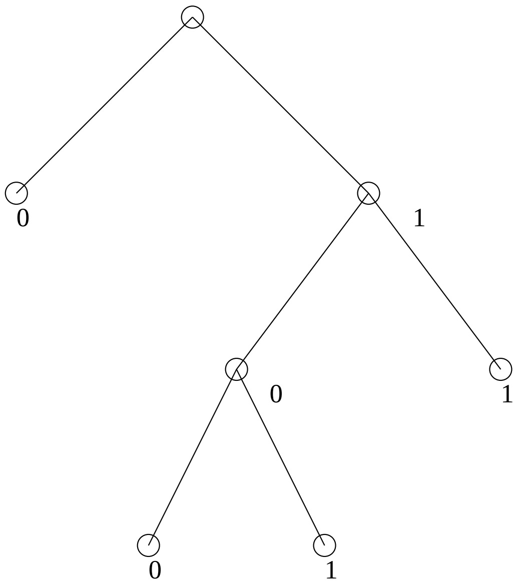

Figure 1: Context tree τ_c for Example 1.

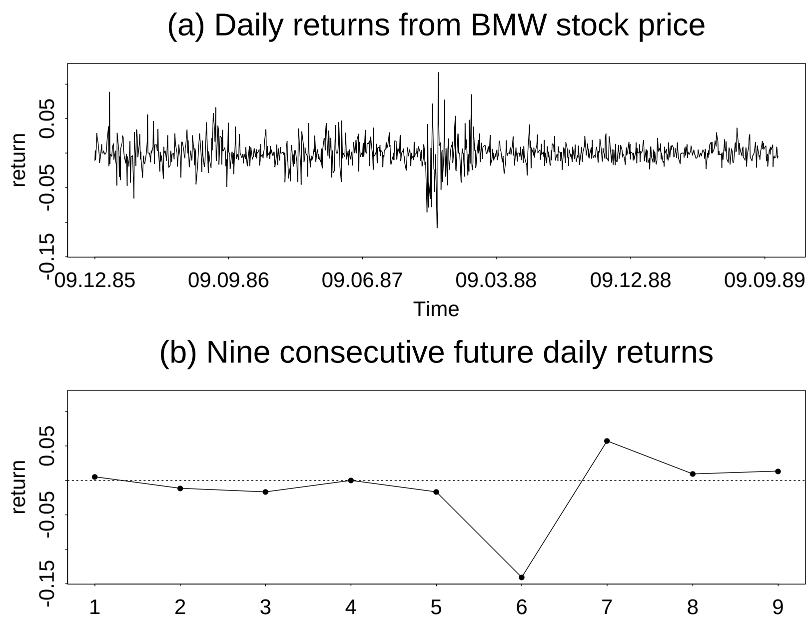

Figure 2: Daily returns from BMW stock price: (a) 1000 days' data, (b) nine consecutive future daily returns.

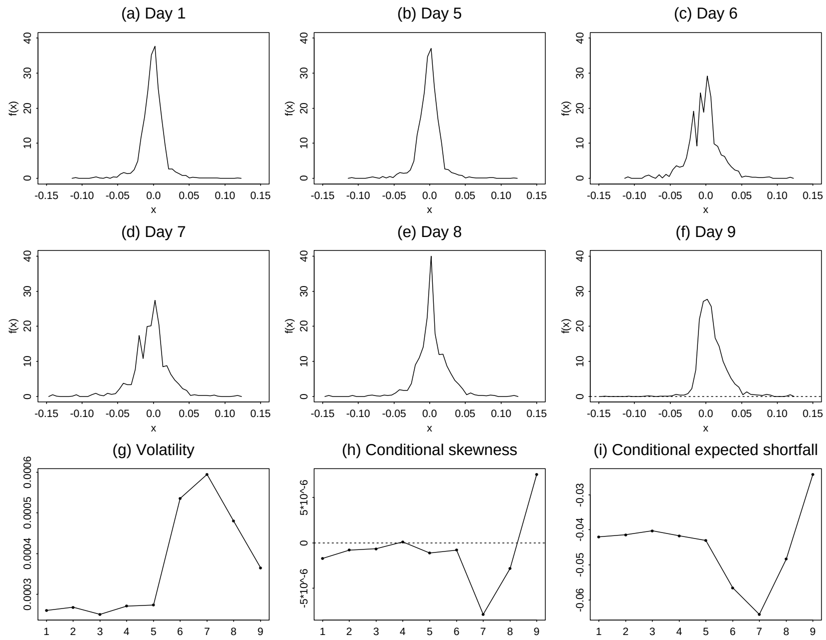

Figure 3: Predictive densities (a)-(f) for six of the nine and summary statistics (g)-(i) for all future value days corresponding to Fig. 2(b).