ETH zürich

ETH Library

Dynamic adaptive partitioning for nonlinear time series

Journal Article

Author(s):
Bühlmann, Peter

Publication date:
1999

Permanent link:
https://doi.org/10.3929/ethz-b-000422552

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
Biometrika 86(3), https://doi.org/10.1093/biomet/86.3.555

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

Biometrika (1999), **86,** 3, *pp.* 555–571
© 1999 Biometrika Trust
Printed in Great Britain

Dynamic adaptive partitioning for nonlinear time series

By PETER BÜHLMANN

Seminar für Statistik, ETH Zürich, CH-8092 Zürich, Switzerland

buhlmann@stat.math.ethz.ch

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; Phi-mixing; Prediction; Quantisation; Stationary process; Tree model.

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 uses 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

$$E_{\text{part}}(Y_t \mid S_{t-1}) = \sum_{j=1}^{J} c_j I(S_{t-1} \in R_j), \qquad (1{\cdot}1)$$

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 $\mathbb{R}^d$ ($d \geqslant 1$), 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 (1999). 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, classification and regression tree, or MARS, multivariate adaptive regression splines method, 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 $E(Y_t | Y_t \leqslant c_{t-1}, Y_{t-1}, Y_{t-2}, \ldots)$ with $c_{t-1} \in \mathbb{R}$ a quantile given the past up to time $t-1$; see an unpublished technical report by A. McNeil and R. Frey from ETH Zürich. 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

$$q : \mathbb{R}^d \rightarrow \mathscr{X} = \{0, 1, \ldots, N-1\} \tag{2·1}$$

a quantiser of $\mathbb{R}^d$ into a categorical set $\mathscr{X} = \{0, 1, \ldots, N-1\}$, inducing a partition

$$\mathbb{R}^d = \bigcup_{x \in \mathscr{X}} I_x, \quad I_x \cap I_y = \varnothing \quad (x \neq y) \tag{2·2}$$

with $y \in I_{q(y)}$ for all $y \in \mathbb{R}^d$.

2·2. *Variable length Markov chains for categorical variables*

Consider a stationary process $\{X_t\}_t$ with values in a finite categorical space $\mathscr{X} = \{0, 1, \ldots, N-1\}$ as in (2·1). We denote by

$$x_i^j = x_j, x_{j-1}, \ldots, x_i \quad (i < j, i, j \in \mathbb{Z} \cup \{-\infty, \infty\})$$

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 § 1 for references.

Definition 1. *Let $\{X_t\}_t$ be a stationary process with values $X_t \in \mathcal{X}$. Denote by $c : \mathcal{X}^{\infty} \to \bigcup_{m=0}^{\infty} \mathcal{X}^m$ a variable projection function which maps $c : x_{-\infty}^{0} \mapsto x_{-l+1}^{0}$, where l is defined by*

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

where $l \equiv 0$ 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 $l(.) = |c(.)|$ determines $c(.)$ and vice versa. The definition of l implicitly reflects the fact that the context length of a variable x_t is $l = |c(x_{-\infty}^{t-1})| = l(x_{-\infty}^{t-1})$, 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*

$$|c(x_{-\infty}^{0})| = l(x_{-\infty}^{0}) \leqslant p$$

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

$$P_c(x_i^j) = \mathrm{pr}(X_i^j = x_i^j), \quad P_c(x_j \mid x_i^{j-1}) = \mathrm{pr}(X_j = x_j \mid X_i^{j-1} = x_i^{j-1}) \quad (i < j)$$

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. Since stationarity is required, a variable length Markov chain is thus completely specified by its transition probabilities $P_c\{x_1 \mid c(x_{-\infty}^{0})\}$, for $x_{-\infty}^{1} \in \mathcal{X}^{\infty}$. 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 $x_{-l(x_{-\infty}^{0})+1}$. 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: $\mathscr{X} = \{0, 1\}$, $p = 3$. 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 left-branching sub-branch represents the symbol 0, and a right-branching sub-branch represents the symbol 1.

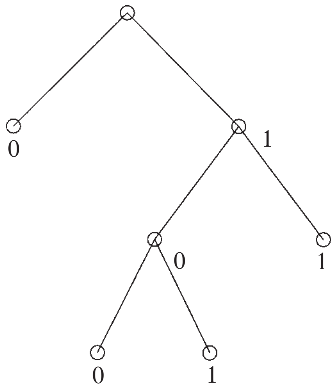

Fig. 1. Context tree τ_c for Example 1.

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*

$$\tau = \tau_c = \{w;\ w = c(x_{-\infty}^{0}),\ x_{-\infty}^{0} \in \mathscr{X}^{\infty}\},$$

$$\tau^T = \tau_c^T = \left\{w;\ w \in \tau_c \text{ and } wu \notin \tau_c \text{ for all } u \in \bigcup_{m=1}^{\infty} \mathscr{X}^m\right\}.$$

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 other than the minimal state space of a variable length Markov chain with context function $c(.)$. An internal node with $b < N = |\mathscr{X}|$ 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, $\mathscr{X}$ and I_x be as in (2·1) and (2·2), respectively. Assume that

(i) $\{X_t\}_t$ is an $\mathscr{X}$-valued variable length Markov chain.

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

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

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

Definition 4. *The process* $\{Y_t\}_t$ *defined by assumptions* (i) *and* (ii) *is called a quantised variable length Markov chain.*

By assumption (ii) a quantised variable length Markov chain has the property that its quantised values, with the correction quantiser q, form a variable length Markov chain, that is $\{q(Y_t)\}_t = \{X_t\}_t$ 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:

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

where '$\leqslant$' 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,

$$\begin{aligned} &q: \mathbb{R}^d \to \mathcal{X}, \quad q(Y_t) = \{q_1(Y_{1,t}), \ldots, q_d(Y_{d,t})\}, \quad Y_t = (Y_{1,t}, \ldots, Y_{d,t}), \\ &q_j: \mathbb{R} \to \mathcal{X}_j = \{0, 1, \ldots, N_j\} \quad (j = 1, \ldots, d), \quad \mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d, \end{aligned} \qquad (2{\cdot}3)$$

with a product space $\mathcal{X}$, labelled arbitrarily by $0, 1, \ldots, N-1$ with $N = N_1 \ldots N_d$. 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 1. The underlying variable length Markov chain $\{X_t\}_t \sim P_c$ on $\mathcal{X}^{\mathbb{Z}}$ satisfies

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

Proposition 1. *Let* $\{Y_t\}_t$ *be a quantised variable length Markov chain as given in Definition* 4, *satisfying Assumption* 1. *Then* $\{Y_t\}_t$ *is ergodic and uniformly mixing with mixing coefficients satisfying* $\phi(i) \leqslant \text{const} \times \rho^i$ *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

$$\pi_{t_1,\ldots,t_m}:(\mathbb{R}^d)^{\mathbb{Z}}\to(\mathbb{R}^d)^m,\quad \pi_{t_1,\ldots,t_m}(y)=y_{t_1},\ldots,y_{t_m}\quad (t_1,\ldots,t_m\in\mathbb{Z}, m\in\mathbb{N})\tag{2·4}$$

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, let '$\Rightarrow$' denote weak convergence.

THEOREM 1. *Let P be a stationary process on* $(\mathbb{R}^d)^{\mathbb{Z}}$ $(d\geqslant 1)$. *Then there exists a sequence* $(P_n)_{n\in\mathbb{N}}$ *of ergodic, $\mathbb{R}^d$-valued quantised variable length Markov chains, such that*

$$P_n\circ\pi^{-1}_{t_1,\ldots,t_m}\Rightarrow P\circ\pi^{-1}_{t_1,\ldots,t_m}\tag{2·5}$$

as $n\to\infty$, for all $t_1,\ldots,t_m\in\mathbb{Z}$, 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

$$f_{Y_{n+m}|Y^n_{-\infty}}(y)=\sum_{x^{n+m}_{n+1}\in\mathcal{X}^m} f_{x_{n+m}}(y)\sum_{j=0}^{m-1} P_c\{x_{n+m-j}\,|\,c(x^{n+m-j-1}_{n+1}X^n_{-\infty})\},\tag{2·6}$$

where

$$x^{n+m-j-1}_{n+1}X^n_{-\infty}=x_{n+m-j-1},x_{n+m-j-2},\ldots,x_{n+1},X_n,X_{n-1},\ldots,X_{-\infty}\quad(j\geqslant 1)$$

and $x^{n+m-j-1}_{n+1}X^n_{-\infty}=X^n_{-\infty}$ 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 $X^n_{-\infty}$ rather than $Y^n_{-\infty}$. 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 squared error m-step-ahead predictor in a quantised variable length Markov chain, it is easy to see that, for a fixed function $g:\mathbb{R}^d\to\mathbb{R}^q$ $(d, q\in\mathbb{N})$,

$$\begin{aligned}E\{g(Y_{n+m})\,|\,Y^n_{-\infty}\}&=E\{g(Y_{n+m})\,|\,c(X^n_{-\infty})\}\\&=\sum_{x^{n+m}_{n+1}\in\mathcal{X}^m}E\{g(Y_{n+m})\,|\,X_{n+m}=x_{n+m}\}\\&\quad\times\sum_{j=0}^{m-1}P_c\{x_{n+m-j}\,|\,c(x^{n+m-j-1}_{n+1}X^n_{-\infty})\}.\end{aligned}\tag{2·7}$$

Again, this predictor takes values only 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 of the self-exciting autoregressive threshold models, SETAR (Tong, 1990, pp. 99–101), 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,

$$c_j = E(Y_t \mid S_{t-1} \in R_j) = E(Y_t \mid V_{t-1} = R_j) = m_p(R_j), \quad m_p : \mathscr{B}^p \rightarrow \mathbb{R}, \tag{2·8}$$

with $\mathscr{B}^p$ 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 nontrivial, 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

$$\begin{gathered} c_j = \sum_{v_t \in R_j} m_1(v_t)\, \mathrm{pr}_{\mathrm{VLMC}}(V_t = v_t \mid V_{t-1} = R_j), \\ m_1(v_t) = m_1(v_{1,t}) = E(Y_t \mid Y_t \in v_{1,t}), \quad v_t = v_{1,t} \times \ldots \times v_{p,t}, \quad m_1 : \mathscr{B}^1 \rightarrow \mathbb{R}, \end{gathered} \tag{2·9}$$

with $\mathscr{B}^1$ the Borel σ-algebra of $\mathbb{R}^1$ and $\mathrm{pr}_{\mathrm{VLMC}}(V_t = v_t \mid V_{t-1} = v_{t-1})$ the probability induced by the variable length Markov chain $\{X_t\}_t$; that is

$$\mathrm{pr}_{\mathrm{VLMC}}(V_t = v_t \mid V_{t-1} = v_{t-1}) = \mathrm{pr}(X_t = x_t \mid c(X_{t-p}^{t-1}) = w_j),$$

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 $\mathscr{B}^1$, 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 $\mathscr{B}^p$ 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 $E(Y_t \mid X_t)$. 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 single 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 $E(Y_t \mid Y_{-\infty}^{t-1})$ 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 for choosing q when $N = |\mathscr{X}| \geqslant 2$ is specified is given by the sample quantiles $\hat{F}^{-1}(.)$ of the data:

$$\hat{q}(y) = \begin{cases} 0 & \text{if } -\infty < y \leqslant \hat{F}^{-1}(1/N), \\ x & \text{if } \hat{F}^{-1}(x/N) < y \leqslant \hat{F}^{-1}\{(x+1)/N\} \quad (x = 1, \ldots, N-2), \\ N-1 & \text{if } \hat{F}^{-1}\{(N-1)/N\} < y < \infty. \end{cases} \tag{3·1}$$

This yields an interval partition of $\mathbb{R}$ with equal numbers of observations per partition cell. Specification of N is discussed in § 3·5. In the multivariate case, we could use a quantiser as in (2·3), 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 adopt the convention that quantities involving time indices $t \notin \{1, \ldots, n\}$ equal zero or are irrelevant. Let

$$N(w) = \sum_{t=1}^{n} I(X_t^{t+|w|-1} = w) \quad (w \in \mathcal{X}^{|w|}) \tag{3·2}$$

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

$$\hat{P}(w) = \frac{N(w)}{n}, \quad \hat{P}(x \mid w) = \frac{N(xw)}{N(w)}, \quad xw = (x_{|x|}, \ldots, x_2, x_1, w_{|w|}, \ldots, w_2, w_1). \tag{3·3}$$

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

$$\Delta_{wu} = \sum_{x \in \mathcal{X}} \hat{P}(x \mid wu) \log\left(\frac{\hat{P}(x \mid wu)}{\hat{P}(x \mid w)}\right) N(wu) \geqslant K$$

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.

CONTEXT ALGORITHM

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 $\tau_{\max}^T$, see Definition* 3, *where $\tau_{\max}^T$ is the biggest tree such that every element, or terminal node, in $\tau_{\max}^T$ has been observed at least twice in the data. Thus $\tau_{\max}^T$ is such that $w \in \tau_{\max}^T$ implies $N(w) \geqslant 2$, and, for every τ^T, where $w \in \tau^T$ implies $N(w) \geqslant 2$, it holds that $\tau^T \preceq \tau_{\max}^T$. 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 $\tau_{(0)}^T = \tau_{\max}^T$.

Step 2. *Examine every element, i.e. terminal node, of $\tau_{(0)}^T$ as follows; the order of examining is irrelevant. Let $c(.)$ be the context function corresponding to $\tau_{(0)}^T$ and let*

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

where wu is an element of $\tau_{(0)}^T$, which we compare with its pruned version $w = x_{-l+2}^0$; if $l = 1$, the pruned version is the empty branch, i.e. the root node.

Prune $wu = x_{-l+1}^0$ to $w = x_{-l+2}^0$ if

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

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 $\tau_{(0)}^T$ yields a possibly smaller tree $\tau_{(1)} \preceq \tau_{(0)}^T$.*

Construct the terminal node context tree $\tau_{(1)}^T$.

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 $P_c\{x_1 \mid c(x^{0}_{-\infty})\}$ by $\hat{P}\{x_1 \mid \hat{c}(x^{0}_{-\infty})\}$, where $\hat{P}(.\mid.)$ 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 Step 2 essentially a log-likelihood-ratio statistic; see Bühlmann & Wyner (1999). Dependence on some values further back in this 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 $c(.)$, that is the tree structured minimal state space τ_c, is discussed in Weinberger et al. (1995) and Bühlmann & Wyner (1999). Asymptotic normality of the transition probabilities $n^{\frac{1}{2}}[\hat{P}\{x_1 \mid \hat{c}(x^{0}_{-\infty})\} - P(x_1 \mid w)]$ 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 2. 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

$$\hat{\Delta}_{\tau_{c_1},\tau_{c_2}}(Y^n_1) = \log\left\{\frac{\hat{f}_{\tau_{c_1}}(Y^n_1)}{\hat{f}_{\tau_{c_2}}(Y^n_1)}\right\},$$

where τ_{c_1}, τ_{c_2} are context trees,

$$\log\{\hat{f}_{\tau_{c_1}}(Y^n_1)\} = \sum_{t=p+1}^{n} \log[\hat{f}\{Y_t \mid c_i(X^{t-1}_{t-p})\}] \quad (i=1, 2)$$

is the loglikelihood of an estimated quantised variable length Markov chain with given $c_i(.)$, induced by τ_{c_i}, and p is the maximal order of $c_1(.)$ and $c_2(.)$. For $i=1, 2$,

$$\hat{f}\{Y_t \mid c_i(X^{t-1}_{t-p})\} = \hat{f}_{X_t}(Y_t)\hat{P}\{X_t \mid c_i(X^{t-1}_{t-p})\}$$

is an estimate in the quantised variable length Markov chain for $f\{Y_t \mid c_i(X^{t-1}_{t-p})\}$ with $\hat{f}_x(.)$ consistent and $\hat{P}(.\mid.)$ as in (3·3).

Assumption 2 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 2, 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* 2 *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 $\{f_x(.);\, x \in \mathscr{X}\}$ can be estimated by some smoothing technique, for example with a kernel estimator

$$\hat{f}_x(y) = \frac{n^{-1}h^{-d}\sum_{t=1}^{n} K\{(y - Y_t)/h\} I(X_t = x)}{n^{-1}N(x)} \quad (y \in \mathbb{R}^d), \tag{3·4}$$

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 $(nh^d)^{\frac{1}{2}}\{\hat{f}_x(y) - f_x(y)\}$ for $x \in \mathscr{X}$ 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 $\text{pr}(Y_t \in E)$ and $\text{pr}(Y_t \in E \mid X_t = x)$ for some measurable set E by

$$\text{est. pr}(Y_t \in E) = n^{-1} \sum_{t=1}^{n} I(Y_t \in E),$$

$$\text{est. pr}(Y_t \in E \mid X_t = x) = \frac{n^{-1}\sum_{t=1}^{n} I(Y_t \in E) I(X_t = x)}{n^{-1}N(x)}.$$

Asymptotic normality of

$$n^{\frac{1}{2}}\{\text{est. pr}(Y_t \in E) - \text{pr}(Y_t \in E)\}, \quad n^{\frac{1}{2}}\{\text{est. pr}(Y_t \in E \mid X_t = x) - \text{pr}(Y_t \in E \mid X_t = x)\}$$

follows from 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 the Context Algorithm in § 3·2. For the predictor in (2·7), we estimate $E\{g(Y_t) \mid X_t = x\}$ by $g(Y)_x^* = N(x)^{-1}\sum_{t=1}^{n} g(Y_t) I(X_t = x)$ and use again the plug-in technique:

$$\hat{E}\{g(Y_{n+m}) \mid Y_1^n\} = \sum_{x_{n+1}^{n+m} \in \mathscr{X}^m} g(Y)_{x_{n+m}}^* \prod_{j=0}^{m-1} \hat{P}\{x_{n+m-j} \mid \hat{c}(x_{n+1}^{n+m-j-1} X_1^n)\}. \tag{3·5}$$

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

$$n^{\frac{1}{2}}[\hat{E}\{g(Y_{n+m}) \mid Y_1^n = y_1^n\} - E\{g(Y_{n+m}) \mid s\}],$$

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

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

where p is the order of the underlying variable length Markov chain. Write

$$\mu_x = E(Y_t \mid X_t = x), \quad \sigma_x^2 = \mathrm{var}(Y_t \mid X_t = x).$$

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

$$l_{\text{quasi}}(\theta;\, Y_1^n) = \sum_{t=p+1}^{n} \log\left[(2\pi\sigma_{X_t}^2)^{-\frac{1}{2}} \exp\left\{-\frac{(Y_t - \mu_{X_t})^2}{2\sigma_{X_t}^2}\right\}\right] + \sum_{t=p+1}^{n} \log[P_c\{X_t \mid c(X_{t-p}^{t-1})\}], \tag{3·6}$$

where $\theta = (\mu_0, \ldots, \mu_{N-1}, \pi)$ with

$$\pi_{wx} = \mathrm{pr}\{X_t = x \mid c(X_{t-p}^{t-1}) = w\} \quad (w \in \tau_c,\ x \in \{0, \ldots, N-2\}).$$

The maximum quasilikelihood estimator for $c(.)$ known,

$$\hat{\theta}_{\text{MQLE}} = \underset{\theta}{\mathrm{argmin}}\, \{-l_{\text{quasi}}(\theta;\, Y_1^n)\}, \tag{3·7}$$

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 $\hat{\theta}_{\text{MQLE}}$ 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 $E(Y_t \mid Y_1^{t-1})$; see McCullagh & Nelder (1980, Ch. 9). In the parametric case in which we find ourselves, a proper AIC-type criterion is of the form

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

In our case, $\dim(\theta) = N + |\tau_c|(N-1)$ with $N = |\mathcal{X}|$. If we replace σ_x^2 with

$$\hat{\sigma}_x^2 = \{N(x) - 1\}^{-1} \sum_{t=1}^{n} (Y_t - \bar{Y}_x)^2 I(X_t = x),$$

our model selection criterion then becomes

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

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 (1999), 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

$$M^2(q, K) = \sum_{t=p+1}^{n} \left\{ \frac{(Y_t - \bar{Y}_{X_t})^2}{\hat{\sigma}^2_{X_t}} + \log(2\pi\hat{\sigma}^2_{X_t}) \right\}$$
$$- 2 \sum_{t=p+1}^{n} \log[\hat{P}\{X_t \mid \hat{c}_K(X_{t-p}^{t-1})\}] + 2\{N + |\tau_{\hat{c}_K}|(N-1)\},$$

where $\hat{c}_K$ is the estimated context function for the $\mathscr{X}$-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 affected only by the term $-2\sum_{t=p+1}^{n} \log[\hat{P}\{X_t \mid \hat{c}_K(X_{t-p}^{t-1})\}]$, thus being exactly the same as when tuning the context algorithm for a categorical valued variable length Markov chain, studied in Bühlmann (1999). For the multivariate model, we choose q and K to minimise.

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

where $\hat{\Sigma}_x = \{N(x) - 1\}^{-1} \sum_{t=1}^{n} (Y_t - \bar{Y}_x)(Y_t - \bar{Y}_x)' I(X_t = x)$ and $\hat{P}(.|.)$ is as in (3·3).

4. NUMERICAL EXAMPLES

4·1. *Simulated data*

We study first the predictive performance of our scheme for simulated univariate data by considering the simple problem of one-step-ahead prediction of observations. 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. Accuracy is measured by

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

with $\hat{Y}_t = \hat{E}(Y_t \mid Y_1^{t-1})$ 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. 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 $K = \chi^2_{N-1;0\cdot95}/2$, $N = |\mathscr{X}|$, 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 $N = |\mathscr{X}|$. 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. We report below parts of a larger simulation study.

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

$$Y_t = \{0\cdot5 + 0\cdot9 \exp(-2\cdot354 Y_{t-1}^2)\} Y_{t-1} - \{0\cdot8 - 1\cdot8 \exp(-2\cdot354 Y_{t-1}^2)\} Y_{t-2} + Z_t, \quad (4\cdot1)$$

with $Z_t \sim \mathcal{N}(0, 0{\cdot}425)$ independently for all t, and Z_t independent from $\{Y_s;\ s < t\}$. The model is specified so that $\operatorname{var}(Y_t) \simeq 1$. Table 1 summarises the results for $n = 4000$ and $n = 5000$. 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'. Projection pursuit autoregression predictably does slightly better in this case where the true model is nonparametric autoregressive with independent, identically distributed innovations. Here MARS is not competitive.

Table 1. *Performances for nonparametric* AR(2) *in* (4·1)

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

QVLMC, quantised variable length Markov chain; AR, minimum AIC linear autoregressive with mean correction; PPreg, projection pursuit on lagged variables with 2 terms; MARS, MARS on lagged variables; oracle, true model.

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

$$Y_t = 0{\cdot}863 \sin(4{\cdot}636 Y_{t-1}) + 0{\cdot}431 \cos(4{\cdot}636 Y_{t-2}) + (0{\cdot}023 + 0{\cdot}5 Y_{t-1}^2)^{\frac{1}{2}} Z_t, \tag{4·2}$$

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 $\operatorname{var}(Y_t) \simeq 1$. Table 2 summarises the results for $n = 4000$; the notation is as in Table 1. 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:

$$Y_{1,t} = 1{\cdot}107 \sin(3{\cdot}629 Y_{1,t-1}) + 0{\cdot}554 \cos(3{\cdot}598 U_{t-1}) + (0{\cdot}038 + 0{\cdot}200 U_{t-1}^2)^{\frac{1}{2}} Z_{1,t},$$

$$U_t = 1{\cdot}107 \sin(3{\cdot}598 U_{t-1}) + 0{\cdot}554 \cos(3{\cdot}629 Y_{1,t-1}) + (0{\cdot}038 + 0{\cdot}200 Y_{1,t-1}^2)^{\frac{1}{2}} Z_{2,t},$$

$$Y_{2,t} = 4{\cdot}721 \left\{ \frac{\exp(U_t)}{1 + \exp(U_t)} - 0{\cdot}5 \right\},$$

Table 2. *Performances for nonparametric* AR(2) *in* (4·2), $n = 4000$

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	—
PPreg, #(lags) = 2	—	0·635	—
MARS, #(lags) = 4	—	0·669	—
Oracle	—	0·523	—

QVLMC, quantised variable length Markov chain; AR, minimum AIC linear autoregressive with mean correction; PPreg, projection pursuit on lagged variables with 2 terms; MARS, MARS on lagged variables; oracle, true model.

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 $\text{var}(Y_{1,t}) \simeq 1$, $\text{var}(Y_{2,t}) \simeq 1$. The results for $n = 4000$ are given in Table 3. 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 attention here to the case $N_1 = N_2$, which is not a necessity. Write

$$\text{PE}_j = 1000^{-1} \sum_{t=n+1}^{n+1000} (\hat{Y}_{j,t} - Y_{j,t})^2 \quad (j = 1, 2), \quad \text{PE}_{\text{tot}} = \text{PE}_1 + \text{PE}_2.$$

Using the best selected quantised variable length Markov chain instead of the linear AR scheme results in a big gain.

Table 3. *Performances for bivariate model*, $n = 4000$

Method	Model dimension	$(\text{PE}_1, \text{PE}_2)$	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	—

QVLMC, quantised variable length Markov chain; AR, minimum AIC linear autoregressive with mean correction.

Other univariate models have also been studied. 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 the 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. Figure 2 shows the data $Y_1, \ldots, Y_n$ and the next nine future values $Y_{n+1}, \ldots, Y_{n+9}$.

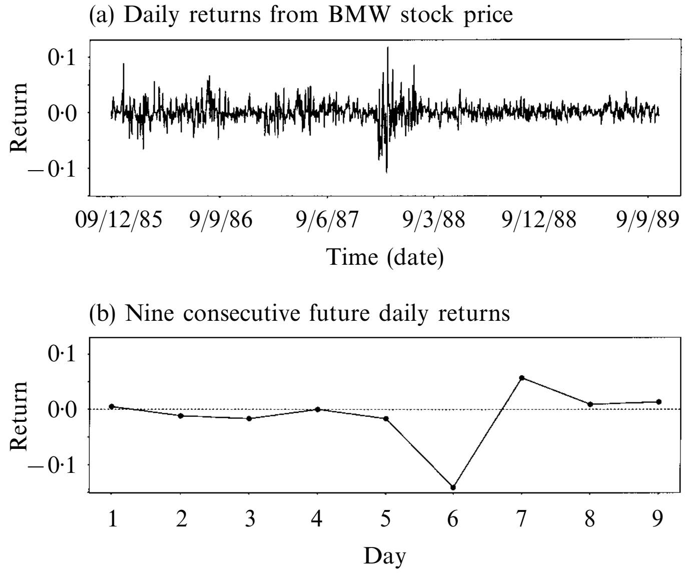

Fig. 2. Daily returns from BMW stock price: (a) 1000 days' data from 9 December 1985 to 9 September 1989, (b) the nine consecutive future daily returns.

The fitted model is with $N = 9$ for the quantiser as in (3·1). Figures 3(a)–(f) show 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 in Fig. 2(b), and Fig. 3(g)–(i) displays the following summary statistics of these predictive densities: the volatilities $\widehat{\text{var}}(Y_t|Y_1^{t-1})$, the skewnesses $\hat{E}[\{Y_t - \hat{E}(Y_t|Y_1^{t-1})\}^3|Y_1^{t-1}]$ and the expected shortfalls $\hat{E}(Y_t|Y_t \leqslant c_{t-1}, Y_1^{t-1})$, with c_{t-1} the estimated conditional 5% quantile of Y_t given Y_1^{t-1}.

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

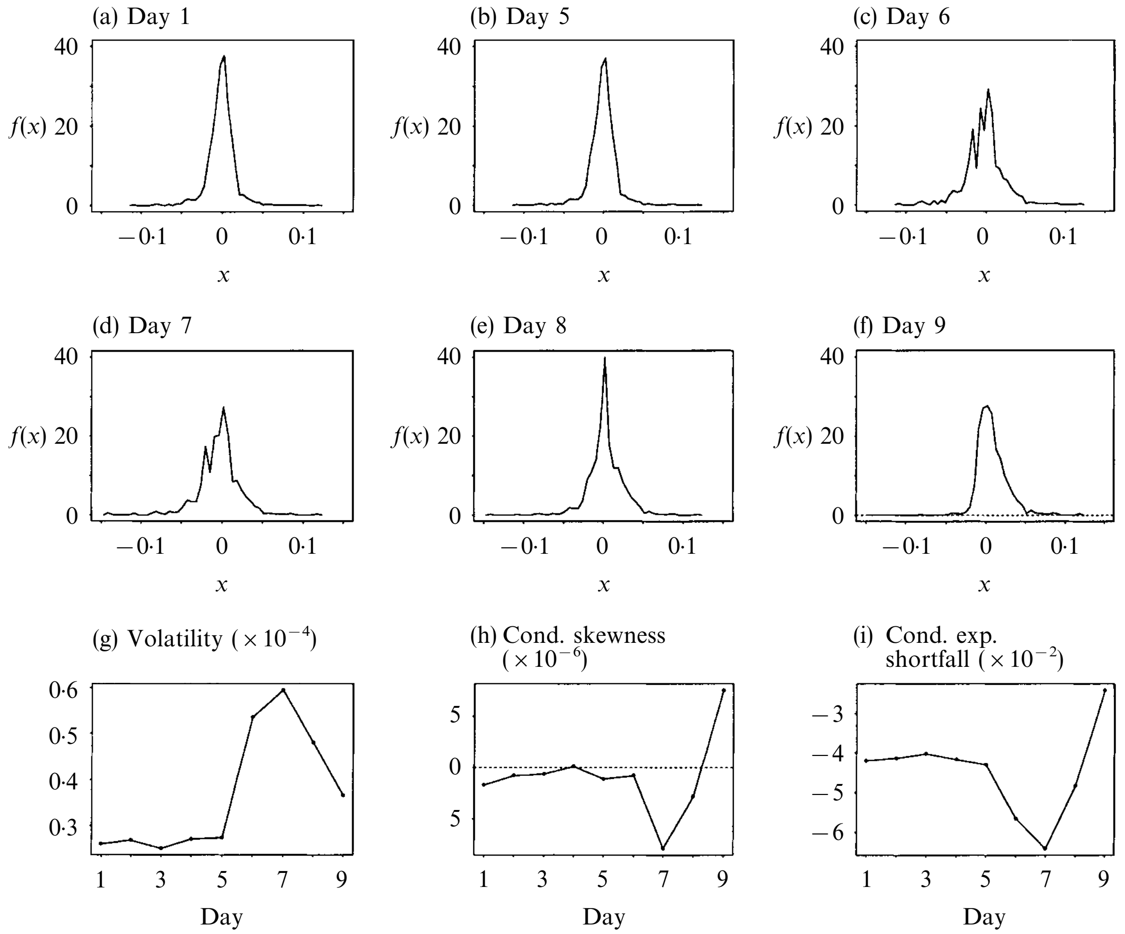

Fig. 3: BMW stock price example. (a)–(f) Predictive densities for six of the nine future days corresponding to Fig. 2(b). (g)–(i) Summary statistics for all nine days, (g) volatility, (h) conditional skewness and (i) conditional expected shortfall.

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.

ACKNOWLEDGEMENT

I thank H. Künsch for interesting discussions. I am also very grateful to the editor for detailed suggestions and two referees for their constructive comments.

APPENDIX

Proofs

Sketch of proof of Theorem 1. The detailed proof can be found in a research report by P. Bühlmann, which is available on the internet at ‘ftp://stat.ethz.ch/Research-Reports/resrep84-Rev.ps.Z’. For notational simplicity we sketch the proof for the univariate case with $d=1$. Let P be a stationary process on $\mathbb{R}^{\mathbb{Z}}$.

In Step 1 we 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. This can be achieved by a partition of $\mathbb{R}$ into intervals which get smaller as $k\rightarrow\infty$, and defining P_k as the probability, with respect to P, of falling into such intervals.

In Step 2 we show that P_k on $\Xi_k^{\mathbb{Z}}$ can be approximated by a sequence of stationary, ergodic Markov chains $(P_{k,l})_{l\in\mathbb{N}}$ on $\Xi_k^{\mathbb{Z}}$. Here $P_{k,l}$ can be constructed as a Markov chain of order $p=p_{k,l}\equiv p_k\rightarrow\infty$ $(k\rightarrow\infty)$. The construction can be based on a transition kernel, which is bounded away from zero for every fixed k and l, and which is close to the corresponding conditional probability with respect to P_k.

In Step 3 we show that $P_{k,l}$ on $\Xi_k^{\mathbb{Z}}$ can be approximated by a sequence $(P_{k,l,m})_{m\in\mathbb{N}}$ of ergodic quantised variable length Markov chains with $P_{k,l,m}$ on $\mathbb{R}^{\mathbb{Z}}$. The construction can be based on

$$\{Y_{t;k,l,m}\}_t=\{Z_{t;k,l}\}_t+\{\varepsilon_{t;m}\}_t\sim P_{k,l,m},$$

with $\{Z_{t;k,l}\}_t\sim P_{k,l}$ and $\varepsilon_{t;m}\sim \mathrm{Un}(-1/(2m),\,1/(2m))$ independently for all t. □

Proof of Proposition 2. We assume $\hat{f}_{X_t}(Y_t)\neq 0$ since $\hat{f}_x(.)$ is consistent. Observe that, by Assumption 2,

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

Therefore,

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

References

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Bühlmann, P. (1999). Model selection for variable length Markov chains and tuning the context algorithm. *Ann. Inst. Statist. Math.* To appear.

Bühlmann, P. & Wyner, A. J. (1999). Variable length Markov chains. *Ann. Statist.* **27**. To appear.

Doukhan, P. (1994). *Mixing. Properties and Examples*, Lecture Notes in Statistics **85**. New York: Springer.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with Discussion). *Ann. Statist.* **19**, 1–141.

Friedman, J. H. & Stuetzle, W. (1981). Projection pursuit regression. *J. Am. Statist. Assoc.* **76**, 817–23.

Gersho, A. & Gray, R. M. (1992). *Vector Quantization and Signal Compression.* Boston, MA: Kluwer.

Lewis, P. A. W. & Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *J. Am. Statist. Assoc.* **86**, 864–77.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

Nobel, A. B. (1997). Recursive partitioning to reduce distortion. *IEEE Trans. Info. Theory* **43**, 1122–33.

Rissanen, J. (1983). A universal data compression system. *IEEE Trans. Info. Theory* **29**, 656–64.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Tong, H. (1990). *Non-linear Time Series. A Dynamical System Approach.* Oxford: Clarendon Press.

Weinberger, M. J., Rissanen, J. & Feder, M. (1995). A universal finite memory source. *IEEE Trans. Info. Theory* **41**, 643–52.

[*Received April* 1998. *Revised December* 1998]