Nonparametric Statistics, Vol. 12, pp. 265–282
Reprints available directly from the publisher
Photocopying permitted by license only

NONPARAMETRIC REGRESSION FOR NONSTATIONARY PROCESSES

CARLO GRILLENZONI*

IUAV, St. Croce 1957, 30135 Venezia, Italy

(Received 1 March 1998; In final form 31 January 1999)

This paper develops recursive kernel estimators for the probability density and the regression function of nonlinear and nonstationary time series. The resulting method is characterized by two smoothing coefficients (the bandwidth and the discounting rate of observations) that may be selected with a prediction error criterion. Statistical properties are investigated under a null hypothesis of stationarity and asymptotic elimination of the discounting. Simulation experiments on complex processes show the ability of the method in estimating time-varying nonlinear regression functions.

Keywords: Nonlinear and nonstationary processes; prediction error criterion; recursive kernel estimators; time-varying regression functions

1. INTRODUCTION

Non-parametric analysis of time series has received increasing attention over the last ten years. Even though it is a relatively small area in the context of nonparametric statistics and time-series analysis, important authors have provided contributions. Original works concern forecasting (Yakowitz, 1985), convergence analysis for dependent sequences (Robinson, 1983) and identification of nonlinear dynamic models (Tjostheim and Auestad, 1994). Rather than a method for modeling and forecasting, nonparametric regression seems to be an important tool for diagnostic and testing non-linearity.

*e-mail: carlog@iuav.unive.it

As pointed out by Priestley (1988) and Tjostheim (1986a), *non-linearity* in time series may be closely related to *non-stationarity*, namely to the time-variability of parameters. This is the typical situation of state-space systems for example. On the other hand, it should be recognized that the most general and realistic situation is represented by processes that are both non-linear and non-stationary. Therefore, the important point is to develop non-parametric methods for estimating regression functions that change over time.

This paper attempts to provide a contribution on this issue. The basic references are the recursive estimators for linear models (*e.g.*, Grillenzoni, 1994) and kernel estimators of the Nadaraya-Watson type. By combining these techniques, a *time-varying non-parametric* regression method for time series is derived. This is characterized by two smoothing coefficients: the well known bandwidth and the discounting rate of observations, which may be selected with a prediction error criterion.

The resulting method is also related to the recursive kernel estimator introduced by Ahmad and Lin (1976). With respect to classical nonparametric smoothers, this approach treats data sequentially ad associate specific bandwidths to each observation. Following the approach of Masry (1987); Roussas and Tran (1992) have derived the asymptotic distribution of the Ahmad-Lin estimator in the case of stationary dependent processes, and conclude that "its full potential has not been appreciated as yet". Finally, a class of recursive kernel estimates based on the Robbins-Monro stochastic approximation scheme, has been developed in engineering by Revesz (1977) and Rutkowsky (1985).

The plan of the work is as follows: Section 2 provides background material for adaptive estimation. Section 3 derives time-varying kernel estimators and investigates their properties under the assumption of stationarity. In Section 4 simulation experiments show the validity of the proposed method in estimating time-varying regression functions.

2. BACKGROUND

The general situation we consider consists of processes of the type

$$Z_t = g_t(Z_{t-1}, Z_{t-2}, \ldots, Z_{t-p}) + a_t, \quad a_t \sim \text{IID}\,(0, \sigma^2) \tag{2.1}$$

with descrete time $t > 0$ and initial conditions $Z_0 = a_0, \ldots, Z_{-p+1} = a_{-p+1}$. The p-variate function $g_t(\cdot)$ is generally non-linear and its structure changes over time. This may occur either with respect to its unknown parameters or its shape. In the first case we speak of *evolution*, whereas in the latter case of *heterogeneity*. The representation (2.1) can be explained and motivated by the following working assumptions:

(A1) The input process $\{a_t\}$ is a sequence of independent and identically distributed (IID) variates having finite variance. We exclude heteroscedasticity because the final goal of a time-varying modeling is to obtain stationary innovations.

(A2) The functions $g_t(\cdot)$ are bounded and differentiable (up to second order) uniformly on R^p and everywhere in $t > 0$. Their boundedness is such that allows the system to be *stochastically stable*, in the sense that $\lim_{z \to \infty} \sup_{t > 0} P(|Z_t| > z) = 0$.

(A3) The output process $\{Z_t\}$ has finite moments (up to fourth order) and is asymptotically independent. It is not ergodic, but may be α, β, ϕ, ρ-*mixing*, see Bosq (1996 p. 15) for the definition of these concepts.

These properties are difficult to check and are only allowed by intrinsic properties of the system functions. In particular, assuming $p = 1$ and using the nonlinear system theory (*e.g.*, Tong, 1990), stochastic stability follows if $g_t(\cdot)$ are *contraction mappings*:

$$\sup_t \sup_z \left(\frac{|g_t(z)|}{|z|} \right) < 1 \tag{2.2}$$

in this case the process (2.1) is bounded in probability (and has second order moments) and is strongly mixing. The proof follows by the fact that in the stationary case, *i.e.*, $g_t(\cdot) = g(\cdot)$, condition (2.2) is sufficient for the *geometric ergodicity* (see Auestad and Tjostheim, 1990 p. 673), a property which implies α-mixing.

In time series literature there are several examples where nonlinear processes may be viewed as time-varying parameter models. The most general class is the doubly stochastic one described by Tjostheim (1986a), which encompasses state-space, bilinear, threshold and other

models. It is obtained from linear processes by modeling the parameters as functions of past events. In the autoregressive case we have

$$Z_t = \phi_1(\mathfrak{I}_{t-1})Z_{t-1} + \cdots + \phi_p(\mathfrak{I}_{t-1})Z_{t-p} + a_t \tag{2.3}$$

where $\mathfrak{I}_{t-1}$ is the space of events (σ-field) generated by $Z_{t-j}, j > 0$. In the case of state-space systems, the parameters $\phi_i(\cdot)$ are linear functions of an exogenous process; if this process coincides with $\{a_t\}$ we have bilinear schemes. In any event, *conditionally* on the space $\mathfrak{I}_{t-1}$, models as (2.3) are linear and have time-varying parameters $\{\phi_{it-1}\}$.

In this context a simple non-parametric estimator for the vector $\underline{\phi}_t = [\phi_{1t-1}, \ldots, \phi_{pt-1}]'$ can be obtained from the technique of *local regression* (see Grillenzoni, 1994). This means discounting observations with, for example, exponential weights

$$\hat{\underline{\phi}}_t(\lambda) = \left(\sum_{i=p+1}^{t} \lambda^{t-i} \underline{X}_i \underline{X}_i' \right)^{-1} \sum_{i=p+1}^{t} \lambda^{t-i} \underline{X}_i Z_i \tag{2.4}$$

where $\underline{X}_i = [Z_{i-1}, \ldots, Z_{i-p}]'$ is the vector of regressors and $0 < \lambda < 1$ is the discounting rate. In this setting, (2.3) may be viewed as a semi-parametric model and (2.4) as its corresponding "one-sided smoother".

Classical non-parametric estimation usually focuses on more complex objects such as the regression function $g_t(z_1, \ldots, z_p) = \mathrm{E}(Z_t | Z_{t-1} = z_1, \ldots, Z_{t-p} = z_p)$, where z_j, $j = 1, \ldots, p$ are auxiliary variables defined on the support of Z_t. If the process (2.1) has order $p = 1$ and data are available in real time, then a sequential estimator of kernel type for $g_t(z)$ is

$$\hat{g}_t(z; h) = \left[\frac{1}{t} \sum_{i=1}^{t} \frac{1}{h} K\left(\frac{z - Z_i}{h} \right) \right]^{-1} \frac{1}{t-1} \sum_{i=2}^{t} \frac{1}{h} K\left(\frac{z - Z_{i-1}}{h} \right) Z_i \tag{2.5}$$
$$\stackrel{\text{def}}{=} \frac{\hat{r}_t(z; h)}{\hat{f}_t(z; h)}$$

where $K(\cdot)$ is the kernel, $h > 0$ is the bandwidth and $\hat{f}_t(z)$ is a sequential estimator of the density of the process. It should be noted that in classical kernel estimation the coefficient $h = h_t$ only depends on the last observation, whereas in the recursive one, it depends on each observation: $h = h_i$, $i = 1, \ldots, t$ (see Roussas and Tran, 1992).

The properties of kernel estimators have been investigated by several authors under conditions of stationarity and mixing (*e.g.*, Robinson, 1983 and Bosq, 1996). However, mixing assumptions are usually difficult to check and are more restrictive than those of ergodicity (*e.g.*, Morvai, Yakowitz and Gyorfi, 1996). For this reason, we consider the following theorem as a basic reference for the paper.

THEOREM 1 *Let* $Z_t = g(Z_{t-1}) + a_t$, $a_t \sim \text{IID}(0, \sigma^2 < \infty)$ *be a strictly stationary, geometrically ergodic process, with moments* $\text{E}(|Z_t|^{2+\eta} | Z_{t-1} = z) < \infty$ *for some* $\eta > 0$. *If the conditions*:

(*B*1) *the functions* $f(a)$, $g(z)$ *are twice continuously differentiable,*
(*B*2) *the variance* $\sigma^2(z) = \text{E}\{[Z_t - g(Z_{t-1} = z)]^2\}$ *is bounded and continuous,*
(*B*3) *the kernel* $K(z)$ *satisfies* $\|K\|_{2+\eta}^2 = \int |K(z)|^{2+\eta} dz < \infty$ *and* $\lim_{|z| \to \infty} [z\, K(z)] = 0$,

then the estimator (2.5) *is such that for any* $z \in R$

$$\sqrt{th}\,[\hat{g}_t(z; h) - g(z)] \xrightarrow{L} \text{N}\left[0; \frac{\sigma^2(z)}{f(z)} \|K\|_2^2\right] \quad \text{as} \quad t \to \infty,\ h \to 0,\ th \to \infty \tag{2.6}$$

Proof See Yakowitz (1985, 1989), Auestad and Tjostheim (1990) and Morvai *et al.* (1996).

In the literature, result (2.6) is usually proved under the condition $th^{5+\eta} \to 0$; however, this mainly serves to minimize the MSE of (2.5). In fact, having $\text{var}(\hat{g}_t) = O(1/\sqrt{th})$ from (2.6) and $\text{bias}(\hat{g}_t) = O(h^2)$ from Auestad and Tjostheim (1990), by letting $h^2 \propto 1/\sqrt{th}$ it follows that the design $h \propto t^{-1/5}$ balances square bias and asymptotic variance.

The technique of local regression belongs to the class of non-parametric estimators (see Hastie and Loader, 1993), but it is relatively heuristic. We establish the relationship between methods (2.4) and (2.5) by deriving the kernel estimator for a model (2.3) with parameters that vary as deterministic functions of the time. In practice, we focus on the *semi-parametric* scheme $Z_t = \sum_{j=1}^{p} \phi_j(t) Z_{t-j} + a_t$ that is stable if $\phi_j(t)$ move inside the parameter space of a stationary AR(p)

process. Given data, the non-parametric part of the model is provided by the functions $\phi_j(\cdot)$ and kernels must be defined for them.

In a first order stationary model, the correspondence arises from the fact that estimator (2.4) minimizes the criterion $Q_t = \sum_{i=2}^{t} \lambda^{t-i} a_i^2$, whereas (2.5) minimizes

$$Q_t(z) = \sum_{i=2}^{t} w_i(z)[Z_i - g(z)]^2 \quad \text{with} \quad w_i(z) = \frac{K[(z - Z_{i-1})/h]}{(t-1)h\hat{f}_t(z)}$$

In our case the regression function is $g(t, z) = \phi(t)z$ and conditionally on $z = Z_{t-1}$ the weights are $w_i(t) = K[(t-i)/h]/(th)$, because $f(t) = 1$. Thus, the resulting kernel estimator becomes (2.4) with weights λ^{t-i} just replaced by $K[(t-i)/h]$. In the following we will focus on the exponential window because it is easier to manage recursively and allows to obtain a suitable expression of the dispersion of the estimator.

3. TIME-VARYING KERNEL ESTIMATION

The sequential implementation is necessary for estimating time-varying parameters, but it is not sufficient because the estimators tend to converge. In order to render the method (2.5) suitable for nonstationary processes, one has to weight observations as in (2.4) so as to retain the tracking ability. The *adaptive* kernel estimator then becomes

$$\begin{aligned} \hat{g}_t(z; h, \lambda) = &\left[\left(\sum_{i=1}^{t} \lambda^{t-i}\right)^{-1} \sum_{i=1}^{t} \frac{\lambda^{t-i}}{h} K\left(\frac{z - Z_i}{h}\right)\right]^{-1} \\ &\left(\sum_{i=2}^{t} \lambda^{t-i}\right)^{-1} \sum_{i=2}^{t} \frac{\lambda^{t-i}}{h} K\left(\frac{z - Z_{i-1}}{h}\right) Z_i \end{aligned} \tag{3.1}$$

Having $(\sum_{i=2}^{t} \lambda^{t-i}) \to 1/(1-\lambda)$ as $t \to \infty$, some terms could be deleted; however, (3.1) is suitable for small samples and its denominator provides the density $\hat{f}_t(z; h, \lambda)$.

Estimator (3.1) is sequential, but every time it processes all available observations. To avoid this drawback one must derive its recursive (on-line) version.

PROPOSITION 1 *As* $t \to \infty$ *the estimator* (3.1) *is equivalent to the recursive algorithm*

$$\hat{f}_t(z) = \lambda \hat{f}_{t-1}(z) + (1-\lambda)\frac{1}{h}K\left(\frac{z - Z_{t-1}}{h}\right) \tag{3.2a}$$

$$\hat{a}_t(z) = [Z_t - \hat{g}_{t-1}(z)] \tag{3.2b}$$

$$\hat{g}_t(z) = \hat{g}_{t-1}(z) + (1-\lambda)\hat{f}_t^{-1}(z)\frac{1}{h}K\left(\frac{z - Z_{t-1}}{h}\right)\hat{a}_t(z) \tag{3.2c}$$

Proof See Appendix A.1.

In (3.2b) we have introduced the prediction error function $\hat{a}_t(z)$ which will have important uses in the following. Focusing on the regression function, the term $(1-\lambda)/h$ is insignificant and could be omitted in (3.2). In this case, however, (3.2a) would no longer provide the recursive estimate of the density. As a comparison with other approaches, one may note that Ahmad and Lin (1976) derived the estimator (3.2c) in the form $\hat{g}_t(z) = \hat{r}_t(z)/\hat{f}_t(z)$, where the numerator was recursively estimated as in (3.2a). Moreover, in the stochastic approximation method of Revesz (1977), the density is not computed, just because the term $(1-\lambda)\hat{f}_t^{-1}(z)$ is replaced by a stepsize coefficient $\alpha > 0$.

Algorithm (3.2) is computationally more efficient than (3.1) and is more transparent in showing its capability of tracking time-varying functions; on the other hand, it requires the initial conditions $f_0(z)$, $g_0(z)$. Under the assumption of stability, the problem of specifying these quantities may be easily solved by setting $Z_0 = a_0$, which implies $f_0(z) = f(a)$ and $g_0(z) = 0$. For $f(a)$ one may assume a non-informative distribution, such as the uniform density with support given by the range of values assigned to z.

Finally, as in standard kernel estimation (*e.g.*, Hardle, 1990), extension of (3.2) to p-th order processes can be obtained by replacing z with the vector $\underline{z} = [z_1, \ldots, z_p]$ and using multivariate kernels. However, it is well known that such functions may be simply obtained as the product of univariate kernels. This leads us to replace the term $h^{-1}K[(z - Z_{t-1})/h]$ with $h^{-p}\Pi_{j=1}^{p}K[(z_j - Z_{t-j})/h]$ in Eqs. (3.2a, c).

Despite the fact that estimator (3.2) is designed for non-stationary processes as (2.1), we now investigate its statistical properties under stationarity. Indeed, this is the only one condition under which a

rigorous inference can be developed. A working condition will be the asymptotic removal of the discounting ($\lambda \to 1$); however, as for the bandwidth in Theorem 1 ($h \to 0$), this must occur at a suitable rate.

PROPOSITION 2 *Under the same assumptions as those in Theorem* 1, *the estimator* (3.2) *is such that*

$$\frac{\sqrt{h}}{\sqrt{1-\lambda}}[\hat{g}_t(z;h,\lambda) - g(z)] \xrightarrow{L} \mathrm{N}\left[0; \frac{1}{2}\frac{\sigma^2(z)}{f(z)}\|K\|_2^2\right] \quad \text{as} \quad t \to \infty,\ h \to 0,\ \lambda \to 1,\ \frac{h}{(1-\lambda)} \to \infty \tag{3.3}$$

Proof See Appendix A.2.

What distinguishes the dispersion in (3.3) from that in (2.6) is the factor 1/2 , which is fundamentally due to the action of the discounting rate. To be more precise, from the expression (A.5) in Appendix we may derive the dispersion as $t \to \infty$

$$\lim_{t\to\infty} \mathrm{E}\{[\hat{g}_t(z) - g(z)]^2\} = \frac{(1-\lambda)}{h(1+\lambda)}\|K\|_2^2\frac{\sigma^2(z)}{f(z)} + o\left(\frac{1-\lambda}{h}\right) \tag{3.4}$$

Clearly, the variance increases as h, $\lambda \to 0$, which means that it is proportional to the adaptive capability of the algorithm (3.2) as realized by the factor $(1-\lambda)/h$.

In kernel estimation it is also customary to evaluate the bias induced by adaptation coefficients. We can show that for stationary processes, only the bandwidth tends to affect the bias of recursive kernel estimates.

PROPOSITION 3 *Under the same assumptions as those in Theorem* 1, *the asymptotic bias of estimator* (3.2) *is*

$$\lim_{t\to\infty} \mathrm{E}[\hat{g}_t(z) - g(z)] = \frac{h^2}{2}\left[g''(z) + 2\frac{g'(z)f'(z)}{f(z)}\right] \times \mu_2(K) + o(h^2) + o\left(\frac{1-\lambda}{h}\right) \tag{3.5}$$

Proof See Appendix A.3.

This result is not surprising because the bias is concerned with first order moments and therefore does not depend on the form of weighting observations. It should be recalled that smoothing coefficients λ, h have very different meanings and roles.

For inferential purposes it is also necessary to have an estimator of the variance $\sigma^2(z) = \mathrm{E}[Z_t - g(Z_{t-1} = z)]^2$ of innovations. In the literature this is usually derived as the difference between kernel estimators of $g_2(z) = \mathrm{E}(Z_t^2 | Z_{t-1} = z)$ and $g^2(z)$ (*e.g.*, Auestad and Tjostheim, 1990). However, Algorithm (3.2) directly estimates the prediction errors $a_t(z) = Z_t - g(Z_{t-1} = z)$, so that a recursive estimator is simply given by

$$\begin{aligned} \hat{\sigma}_t^2(z) &= (1-\lambda)\sum_{i=2}^{t} \lambda^{t-i}[Z_t - \hat{g}_{t-1}(z)]^2 \\ &= \lambda\,\hat{\sigma}_{t-1}^2(z) + (1-\lambda)\hat{a}_t^2(z) \end{aligned} \tag{3.6}$$

Also in this case, the condition $Z_0 = a_0$ provides a suitable starting value: $\sigma_0^2(z) = \sigma^2$.

At this point one may wonder what the meaning of the above framework is, because Algorithm (3.2) is designed for time-varying systems, but its distribution (3.3) is obtained under the assumption of stationarity. The answer is that (3.3) can be used in tests for statistical significance and constancy of regression functions. In fact, in these cases the null hypotheses involve constant parameters, and therefore stationarity. For example, when testing for $\mathrm{H}_0 : g_{t_0}(z_0) = 0$ (which means that Z_t is white noise at time t_0 and conditionally on $Z_{t-1} = z_0$) one can use the asymptotic confidence interval

$$\mathrm{P}\left[g_{t_0}(z_0) \in \left(\hat{g}_{t_0}(z_0) \pm 1.96\left[\frac{(1-\lambda)}{h(1+\lambda)}\|K\|_2^2 \frac{\hat{\sigma}_{t_0}^2(z_0)}{\hat{f}_{t_0}(z_0)}\right]^{1/2}\right)\right] \approx 0.95$$

On the other hand, when testing the hypothesis of time-constancy $\mathrm{H}_0 : g_{t_1}(z_0) = g_{t_2}(z_0)$, two confidence intervals centered on $\hat{g}_{t_1}(z_0)$, $\hat{g}_{t_2}(z_0)$ must be constructed.

OPTIMAL DESIGN From (3.4) it is clear that the design of the coefficients h, λ should provide a suitable trade-off between adaptive

capability and variability of estimates. In absence of *a-priori* information on the path of the regression function, a data-driven approach must be used for their selection. A suitable criterion is based on the prediction errors (3.2b) evaluated at the empirical points $z = Z_{t-1}$. As in Grillenzoni (1994), this involves minimizing a quadratic functional over the available sample $\{Z_t, t = 1, \ldots, T\}$

$$\hat{h}_T, \hat{\lambda}_T = \arg\min Q_T(h, \lambda) = \left[\sum_{t=2}^{T} \hat{a}_t^2(z = Z_{t-1}) \right] \tag{3.7}$$

where the algorithm for computing $Q_T(\cdot)$ is provided by (3.2) itself. Having $\hat{a}_t(z) = a_t - [\hat{g}_{t-1}(z) - g(z)]$, it is clear that solutions of (3.7) tend to optimize both model fitting and MSE accuracy of the recursive estimates. In general, it may be viewed as a sequential cross-validation procedure in which the omitted observation is always the last.

If minimization (3.7) is carried out iteratively, then resulting estimates belong to the class of conditional least squares (CLS), analyzed by Hall and Heyde (1980) and Tjostheim (1986b). Following this analysis, one can conclude that under the assumptions of Theorem 1 and suitable smoothness conditions for the loss function, such as:

(C1) $T^{-1}\partial Q_T/\partial h$, $T^{-1}\partial Q_T/\partial \lambda$ converge to zero with probability one (w.p.1), as $T \to \infty$.

(C2) $\underline{M}_T = T^{-1}\partial^2 Q_T/\partial h\ \partial \lambda$ converges w.p.1 to a positive definite matrix, as $T \to \infty$.

(C3) $\lim_{T\to\infty} \sup_{\varepsilon \to 0} \varepsilon^{-1} |\underline{M}_T(h^*, \lambda^*) - \underline{M}_T(h, \lambda)| < \infty$, w.p.1 for $|(h^*, \lambda^*) - (h, \lambda)| < \varepsilon$,

the estimates $\hat{h}_T$, $\hat{\lambda}_T$ in (3.7) converge w.p.1 to the optimal values h^*, λ^*. Under additional, more restrictive, assumptions it is also possible to prove the asymptotic normality.

These results are significantly related to the properties of cross-validation estimates of the bandwidth in cross sectional data. For example, Hardle, Hall and Marron (1988) proved weak consistency and asymptotic normality, even though for finite samples the approximate distribution is related to a chi-square (Chiu, 1990). On the other hand, in Tjostheim (1986b) the consistency of CLS estimates has been extended to processes which depend on initial conditions. This means that it might be obtained even for non-stationary processes as (2.1) under assumptions (A) and (C).

4. SIMULATION EXPERIMENTS

Numerical simulations were performed to check the adaptive capability of the non-parametric framework (3.2)–(3.7). In particular, we were interested in checking if Algorithm (3.2) can estimate time-varying non-linear regression functions, and if criterion (3.7) is able to provide suitable values for the adaptation coefficients h, λ. All computations were carried out with the MATLAB 4.0 package on a personal computer.

We considered complex first order autoregressive processes of the type

$$Z_t = \sin(t/16)\sin(Z_{t-1}) + a_t, \quad a_t \sim \mathrm{IU}(-1,+1) \tag{4.1a}$$

$$Z_t = \sin(t/16)\log(|Z_{t-1}|) + a_t, \quad a_t \sim \mathrm{IU}(-1,+1) \tag{4.1b}$$

$$Z_t = \sin(t/16)\exp(-|Z_{t-1}|) + a_t, \quad a_t \sim \mathrm{IU}(-1,+1) \tag{4.1c}$$

where common features are that inputs a_t have independent uniform (IU) densities with support $(-1,+1)$ and the parameter $\phi(t)$ is a sinusoidal function with period 100. The regression functions $\sin(z)$, $\log(|z|)$, $\exp(-|z|)$ are non-linear, but seem relatively simple. In reality, the combination with $\phi(t)$ makes the global pattern of $g_t(z)$ complex. In particular, at time $t = 50$ their paths change sign (see Figs. 2a and 3a).

For processes (4.1), $N = 100$ realizations of sample size $T = 100$ were generated with the initial condition $Z_0 = a_0$. Although the parameter $\phi(t)$ meets the unit circle at $t = 25, 75$ we did not encountered problems of instability. Mean values and standard errors of the CLS estimates of the adaptation coefficients are reported in Table I. We may see that all estimates belong to the admissible range $0 < h, \lambda < 1$ and the reduction of the predictive statistic Q_T is significant (especially in the model (4.1b)).

Figure 1(a, b) shows the kernel estimates of the densities of coefficients $\{\hat{\lambda}_k, \hat{h}_k\}_1^N$ in the first simulation. Both were obtained with band-

TABLE I Mean values and standard errors of estimates (3.7) applied to processes (4.1)

Model	$\bar{h}_N$	$SE(\hat{h}_k)$	$\bar{\lambda}_N$	$SE(\hat{\lambda}_k)$	$\bar{Q}_T(\hat{a}_t)$	$\bar{Q}_T(Z_t)$	h^*_{MISE}	λ^*_{MISE}
(4.1a)	.392	.094	.891	.035	42.6	53.1	.30	.75
(4.1b)	.206	.082	.813	.111	74.4	116.	.10	.70
(4.1c)	.401	.129	.882	.037	42.8	52.8	.25	.75

* Designates coefficients that minimize the integrated MSE, as obtained with a search procedure.

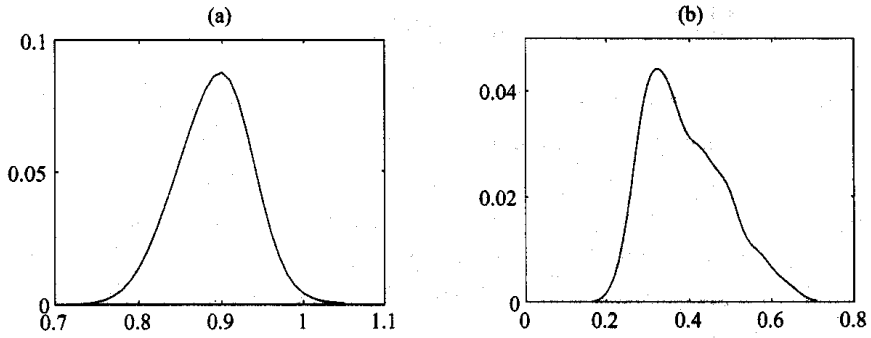

FIGURE 1 Kernel density estimates of estimates (3.7): (a) Factor $\hat{\lambda}_k$, (b) Bandwidth $\hat{h}_k$.

width $h = .03$. Note that the distribution of $\hat{\lambda}_k$ is approximately normal, whereas that of $\hat{h}_k$ tends to be asymmetric. This confirms that the convergence in distribution of bandwidth estimates is slow.

Figures 2 and 3 show main graphical aspects of numerical simulations with models (4.1a) and (4.1b). Those concerned with (4.1c) are omitted

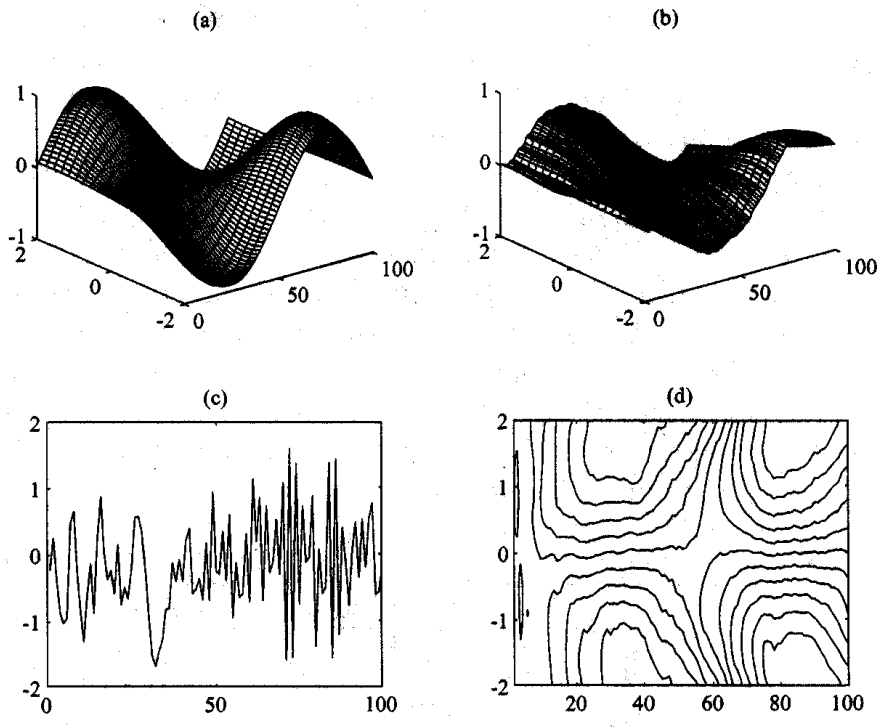

FIGURE 2 Graphical aspects of simulation with model (4.1a): (a) Theoretical regression function $g_t(z)$, (b) mean value of estimates $\bar{g}_t(z)$, (c) A typical realization of Z_t, (d) Contour of $\bar{g}_t(z)$, (e) Sections of $\bar{g}_t(z)$ for $-2 \leq z \leq +2$, (f) Sections of $\bar{g}_t(z)$ for $1 \leq t \leq 100$.

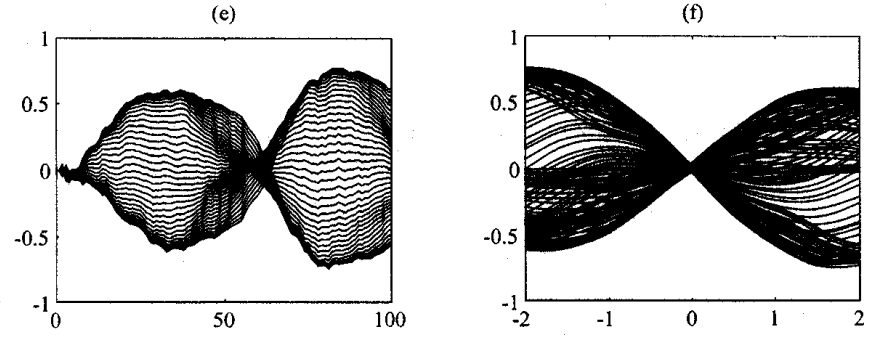

FIGURE 2 (Continued).

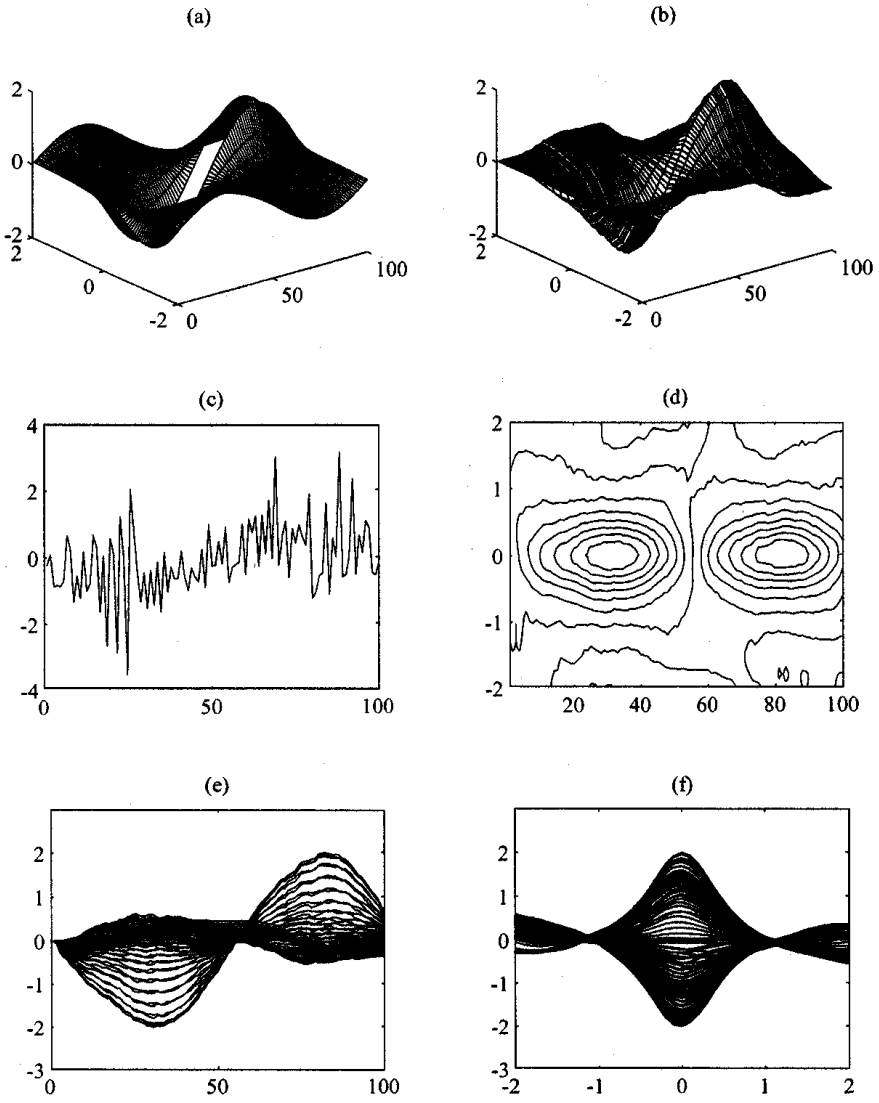

FIGURE 3 Graphical aspects of simulation with model (4.1b): (a) Theoretical regression function $g_t(z)$, (b) mean value of estimates $\bar{g}_t(z)$, (c) A typical realization of Z_t, (d) Contour of $\bar{g}_t(z)$, (e) Sections of $\bar{g}_t(z)$ for $-2 \leq z \leq +2$, (f) Sections of $\bar{g}_t(z)$ for $1 \leq t \leq 100$.

because they are similar to those of (4.1b). Computations were based on the estimates underlying the values in Table I. In particular, given the k-th realization Z_{kt} and the coefficients $\hat{h}_k, \hat{\lambda}_k$ obtained with (3.7), the functions $\hat{g}_{kt}(z_j)$ were generated with Algorithm (3.2) for a grid of values $-2 \leq z_j \leq +2$; finally, their mean value $\bar{g}_t(z_j)$ was computed over k.

To comment on these results, we can state that the capability of framework (3.2)–(3.7) to estimate regression functions that change smoothly over time is satisfactory. First, we may note that global pattern of $\bar{g}_t(z)$ clearly resemble those of $g_t(z)$ (see Figs. 2b and 3b). Second, the mean values $\bar{h}_N, \bar{\lambda}_N$ in Table I are close to the optimal MSE coefficients and allow a significant reduction of the predictive statistic Q_T.

References

Ahmad, I. A. and Lin, P. (1976) Nonparametric sequential estimation of a multiple regression function. *Bulletin of Mathematical Statistics*, **17**, 63–75.

Auestad, B. and Tjostheim, D. (1990) Identification of nonlinear time series: First order characterization and order determination. *Biometrika*, **77**, 669–687.

Bosq, D. (1996) *Nonparametric Statistics for Stochastic Processes*, Springer, New York.

Chiu, S.-T. (1990) On the asymptotic distributions of bandwidth estimates. *The Annals of Statistics*, **18**, 1696–1711.

Grillenzoni, C. (1994) Optimal recursive estimation of dynamic models. *Journal of the American Statistical Association*, **89**, 777–787.

Grillenzoni, C. (1996) Testing for causality in real time. *Journal of Econometrics*, **73**, 355–377.

Hall, P. and Heyde, C. C. (1980) *Martingale Limit Theory and its Applications*, Academic Press, New York.

Hardle, W. (1990) *Applied Nonparametric Regression.* Cambridge University Press, Cambridge.

Hardle, W., Hall, P. and Marron, S. (1980) How far are automatically chosen smoothing parameters from their optimum? *Journal of American Statistical Association*, **83**, 86–101.

Hastie, T. and Loader, C. (1993) Local regression: automatic kernel carpentry. *Statistical Science*, **8**, 120–143.

Lindoff, B. and Holst, J. (1996) Bias and covariance of the RLS estimator with exponential forgetting in vector autoregressions. *Journal of Time Series Analysis*, **17**, 553–570.

Masry, E. (1987) Almost sure convergence of recursive density estimation for weakly dependent processes. *Statistics and Probability Letters*, **5**, 249–254.

Morvai, G., Yakowitz, S. and Gyorfi, L. (1996) Nonparametric inference for ergodic stationary time series. *Annals of Statistics*, **24**, 370–379.

Priestley, M. B. (1988) *Nonlinear and Nonstationary Time Series Analysis*, Academic Press, London.

Revesz, P. (1977) How to apply the method of stochastic approximation in the nonparametric estimation of regression functions. *Mathematische Operationsforschung, Series Statistics*, **8**, 119–126.

Robinson, P. M. (1983) Non-parametric estimation for time series models. *Journal of Time Series Analysis*, **4**, 185–208.

Roussas, G. G. and Tran, L. T. (1992) Asymptotio normality of the recursive kernel regression estimate under dependence conditions. *Annals of Statistics*, **20**, 98–120.
Rutkowski, L. (1985) Real-time identification of time-varying systems by nonparametric algorithms. *International Journal of Systems Science*, **16**, 1123–1130.
Tong, H. (1990) *Time Series Analysis: a Dynamical Systems Approach*. Clarendon Press, Oxford.
Tjostheim, D. (1986a) Some doubly stochastic time series models. *Journal of Time Series Analysis*, **7**, 51–72.
Tjostheim, D. (1986b) Estimation in nonlinear time series models. *Stochastic Processes and their Applications*, **23**, 251–273.
Tjostheim, D. and Auestad, B. (1994) Non-parametric identification of non-linear time series. *Journal of American Statistical Association*, **89**, 1398–1419.
Yakowitz, S. J. (1985) Nonparametric density estimation, prediction and regression for markov sequences. *Journal of American Statistical Association*, **80**, 215–221.
Yakowitz, S. J. (1989) Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *Journal of Multivariate Analysis*, **30**, 124–137.

A. APPENDIX: TECHNICAL DETAILS

A.1. Proof of Proposition 1

For t sufficiently large, we can introduce in (3.1) the approximation $\left(\sum_i^t \lambda^{t-i}\right)^{-1} \approx (1-\lambda)$ and we may compute the estimated density on lagged values of Z_t, namely $\hat{f}_t(z) = h^{-1}(1-\lambda)\sum_{i=2}^t \lambda^{t-i} K[(z-Z_{i-1})/h]$. Now, using the notations $\hat{f}_t = \hat{f}_t(z;h,\lambda), \hat{g}_t = \hat{g}_t(z;h,\lambda)$ and $K_i = K[(z-Z_i)/h]$ we have

$$\begin{aligned}\hat{f}_t &= \frac{(1-\lambda)}{h}\left(\lambda\sum_{i=2}^{t-1}\lambda^{t-1-i}K_{i-1} + K_{t-1}\right)\\ &= \lambda\hat{f}_{t-1} + \frac{(1-\lambda)}{h}K_{t-1}\end{aligned} \tag{A.1}$$

and

$$\begin{aligned}\hat{g}_t &= \hat{f}_t^{-1}\left[\lambda\sum_{i=2}^{t-1}\lambda^{t-1-i}K_{i-1}Z_i + K_{t-1}Z_t\right]\frac{(1-\lambda)}{h}\\ &= \hat{f}_t^{-1}\left[\lambda\hat{f}_{t-1}\hat{g}_{t-1} + K_{t-1}Z_t\frac{(1-\lambda)}{h}\right]\\ &= \hat{f}_t^{-1}\left[\hat{f}_t\hat{g}_{t-1} + \frac{(1-\lambda)}{h}K_{t-1}(Z_t - \hat{g}_{t-1})\right]\\ &= \hat{g}_{t-1} + \frac{(1-\lambda)}{h}\hat{f}_t^{-1}K_{t-1}\hat{a}_t\end{aligned} \tag{A.2}$$

which clearly provides (3.2) with the assumed notations.

A.2. Proof of Proposition 2

From Proposition 1 the estimator (3.2) is equivalent to (3.1), which converges in probability to (2.5) as $\lambda \to 1$. This means that the estimates $\hat{f}_t(z)$, $\hat{g}_t(z)$ are consistent and asymptotically normal as $\lambda \to 1$, $h \to 0$, $th \to \infty$. In Grillenzoni (1996) it is shown that the variance of exponentially weighted statistics is of order $O(1-\lambda, 1/t)$, therefore we may write

$$[\hat{f}_t(z;h,\lambda) - f(z)] = [\hat{g}_t(z;h,\lambda) - g(z)] = O_p(\sqrt{1-\lambda}; \sqrt{1/th}) \quad \text{(A.3)}$$

Now subtracting $g(z)$ from both sides of (A.3) and multiplying by $\hat{f}_t(z)$ we have

$$\hat{f}_t(\hat{g}_t - g) = \hat{f}_t(\hat{g}_{t-1} - g) + \frac{(1-\lambda)}{h} K_{t-1}(Z_t - \hat{g}_{t-1})$$

and using expression (A.1) and $Z_t = g(Z_{t-1} = z) + a_t(z)$, the above becomes

$$\hat{f}_t(\hat{g}_t - g) = \left[\lambda \hat{f}_{t-1} + \frac{(1-\lambda)}{h} K_{t-1}\right](\hat{g}_{t-1} - g) + \frac{(1-\lambda)}{h} K_{t-1}[a_t - (\hat{g}_{t-1} - g)]$$

which simplifies as

$$\hat{f}_t(\hat{g}_t - g) = \lambda \hat{f}_{t-1}(\hat{g}_{t-1} - g) + \frac{(1-\lambda)}{h} K_{t-1} a_t$$

Since $\lambda < 1$, this provides a stable difference equation whose solution is

$$\hat{f}_t(\hat{g}_t - g) = \sum_{i=2}^{t} \lambda^{t-i} \frac{(1-\lambda)}{h} K_{i-1} a_i$$

From (A.3) a similar result holds by replacing $\hat{f}_t(z)$ by $f(z)$ and adding a term $o_p(1)$. Squarring the resulting expression we find that

$$\frac{h}{(1-\lambda)}(\hat{g}_t - g)^2 = \frac{(1-\lambda)}{h f^2} \sum_{i=2}^{t} \sum_{j=2}^{t} \lambda^{2t-i-j} K_{i-1} a_i a_j K_{j-1} + o_p(1) \quad \text{(A.4)}$$

Since Z_{t-1} and a_t are orthogonal, for $i > j$ and $\mathfrak{I}_i = \text{set}(Z_i, Z_{i-1} \ldots)$ we have

$$\begin{aligned} \mathrm{E}(K_{i-1}a_i a_j K_{j-1}) &= \mathrm{E}\left[\mathrm{E}(K_{i-1}a_i a_j K_{j-1}|\mathfrak{I}_{i-1})\right] \\ &= \mathrm{E}\left[K_{i-1}a_j K_{j-1}\mathrm{E}\left(a_i|\mathfrak{I}_{i-1}\right)\right] = 0 \end{aligned}$$

and the same holds for $i < j$. Therefore, taking expectation in (A.4) we obtain

$$\frac{h}{(1-\lambda)}\mathrm{E}\left[(\hat{g}_t - g)^2\right] = \frac{(1-\lambda)}{h f^2}\sum_{i=2}^{t}\lambda^{2(t-i)}\mathrm{E}\left(K_{i-1}^2\right)\mathrm{E}(a_i^2) + o(1)$$

and under the assumption of stationarity for Z_t, which implies that of K_t, we also have

$$\lim_{t\to\infty}\frac{h}{(1-\lambda)}\mathrm{E}\left[(\hat{g}_t - g)^2\right] = \frac{(1-\lambda)}{h f^2}\frac{1}{(1-\lambda^2)}\mathrm{E}\left(K_{t-1}^2\right)\mathrm{E}(a_t^2) + o(1) \tag{A.5}$$

Finally, the dispersion in (3.3) can be obtained from $(1-\lambda^2) = (1-\lambda)(1+\lambda)$ and the well known result $\mathrm{E}(K_{t-1}^2) = \mathrm{E}[K^2((z - Z_{t-1})/h)] = h f(z)\|K\|_2^2 + o(h)$ (see Hardle, 1990).

A.3. Proof of Proposition 3

Evaluation of bias involve first order moments. In order to exploit the linearity property of the operator $\mathrm{E}(\cdot)$, a linearization of $\hat{g}_t(z)$ is investigated. As in Auestad and Tjostheim (1990) we consider a Taylor expansion of $\hat{g}_t$ in $g = r/f$ of the form

$$\begin{aligned} (\hat{g}_t - g) = \frac{(\hat{r}_t - r)}{f} - \frac{r(\hat{f}_t - f)}{f^2} - \frac{(\hat{r}_t - r)(\hat{f}_t - f)}{f^2} + \frac{r(\hat{f}_t - f)^2}{f^3} \\ + o_p\left(\frac{\sqrt{1-\lambda}}{\sqrt{h}}; \frac{1}{\sqrt{t}}\right) \end{aligned} \tag{A.6}$$

the remainder term depends on the rate of convergence of estimates. Now, as for (3.4) we can show that

$$\lim_{t\to\infty} \mathrm{E}[(\hat{f}_t - f)^2] = \frac{(1-\lambda)}{h(1+\lambda)} \|K\|_2^2 f(z) + o\left(\frac{1-\lambda}{h}\right)$$

$$\lim_{t\to\infty} \mathrm{E}[(\hat{r}_t - r)(\hat{f}_t - f)] = \frac{(1-\lambda)}{h(1+\lambda)} \|K\|_2^2 f(z)g(z) + o\left(\frac{1-\lambda}{h}\right)$$

moreover, it is well known that

$$\mathrm{E}(\hat{r}_t - r) = \frac{h^2}{2} r''(z)\mu_2(K) + o(h^2),$$

$$\mathrm{E}(\hat{f}_t - f) = \frac{h^2}{2} f''(z)\mu_2(K) + o(h^2)$$

where $\mu_2(K) = \int z^2 K(z)dz$ (see Hardle, 1990). Finally, the bias (3.5) can be easily obtained by computing $r'' = ((gf)')'$ and substituting the above expressions into (A.6).