Vidar Hjellvik, Qiwei Yao and Dag Tjøstheim

Linearity testing using local polynominal approximation

Article (Accepted version) (Refereed)

Original citation:
Hjellvik, Vidar and Yao, Qiwei and Tjostheim, Dag (1998) Linearity testing using local polynominal approximation. Journal of statistical planning and inference, 68 (2). pp. 295-321. DOI: 10.1016/S0378-3758(97)00146-8

Linearity Testing using Local Polynomial Approximation

by

Vidar Hjellvik
Department of Mathematics
University of Bergen
5007 Bergen
NORWAY

Qiwei Yao
Institute of Mathematics and Statistics
University of Kent at Canterbury
Canterbury, Kent CT2 7NF
United Kingdom

Dag Tjøstheim
Department of Mathematics
University of Bergen
5007 Bergen
NORWAY

Abstract

We use local polynomial approximation to estimate the conditional mean and conditional variance, and test linearity by using a functional measuring the deviation between the nonparametric estimates and the parametric estimates based on a linear model. We also employ first and second order derivatives for this purpose, and we point out some advantages of using local polynomial approximation as opposed to kernel estimation in the context of linearity testing. The asymptotic theory of the test functionals is developed in some detail for a special case. It is used to draw qualitative conclusions concerning the bandwidth, but in order to apply the asymptotic distribution to specific testing problems very large sample sizes are needed. For moderate sample sizes we have examined a bootstrap alternative in a large variety of situations. We have tried bandwidths suggested by asymptotic results as well as bandwidths obtained by cross-validation.

1 Introduction

Recently Hjellvik and Tjøstheim (1995,1996) have derived linearity tests based on nonparametric estimates of the conditional mean and the conditional variance. A more general problem of this type was considered by Härdle and Mammen (1993). In all of these cases the estimation was carried out using kernel (Nadarya-Watson) type estimates.

Local polynomial estimation is an alternative to the kernel method. It has been promoted in particular by Fan (1992, 1993), and it has been applied for example to study the interface between nonparametrics and chaos (Yao and Tong 1994, and Fan, Yao and Tong 1996). In this paper we examine its potential in linearity testing. For example it is convenient to look at derivatives of nonparametric estimates in this framework, and one can construct new tests of linearity exploiting that the first order derivative is a constant, and the second order derivative is zero for a linear model. It is also easier to look at the transition between parametric and nonparametric modeling. This transition is intimately connected to the choice of bandwidth. Choosing the bandwidth is an important aspect of nonparametric linearity testing, but it was virtually neglected in Hjellvik and Tjøstheim (1995,1996). In the present paper it is studied in some detail and both data driven and theoretically determined bandwidths are investigated.

In contrast to Hjellvik and Tjøstheim (1995,1996) we have worked out a fair amount of asymptotic theory. One reason for this is that the asymptotic theory yields useful input to the problem of choosing the bandwidth. Also the asymptotic theory is of interest in itself, and in Lemma 3.2 we extend some results on degenerate U-statistics, which has hitherto only been proved for independent and identically distributed (*iid*) random variables. The details of the derivations are quite technical and due to considerations of length they had to be omitted in the present version. They are included in Hjellvik *et al.* (1996), however. Again it is found that very large sample sizes are needed to obtain a good approximation to the asymptotic distribution of the test functionals. For moderate and small sample sizes a much better approximation is achieved by bootstrapping. We present a number of examples, both simulated and real, to illustrate our procedures. We also discuss briefly the interpretation of very low p-values.

2 Preliminaries

Suppose that $\{X_t, Y_t\}$ is a strictly stationary discrete-time stochastic process with $X_t \in R^d$ and Y_t one-dimensional. Let $p(\cdot)$ denote the smooth density function of X_t. Given observations

$\{(X_t, Y_t)\ 1 \leq t \leq n\}$, we are basically interested in testing whether the conditional expectation $m(x) = \mathrm{E}\{Y_t | X_t = x\}$ is a linear function. We write

$$Y_t = m(X_t) + \epsilon_t, \quad t \geq 1, \tag{2.1}$$

where $\mathrm{E}\{\epsilon_t | X_t\} = 0$ for all t. This setup includes the autoregressive model as a special case in which X_t consists of some lagged variables of Y_t. We do not assume that $\{\epsilon_t, t \geq 1\}$ are *iid* . This, in particular, allows for the case of conditional heteroscedasticity.

Assume that $\mathrm{E}Y_t = 0$ and $\mathrm{E}X_t = 0$. Our hypothesis can be specified as

$$H_0 : m(\cdot) \text{ is linear, i.e. } m(x) = x^\tau \theta,\ \theta \in R^d \text{ unknown,}$$

where τ denotes the transpose, against

$$H_1 : m(\cdot) \text{ is nonlinear.}$$

The curse of dimensionality means that it is difficult to estimate $m(\cdot)$ nonparametrically unless d is small, and we have chosen (cf. Hjellvik and Tjøstheim 1995,1996) to use the one dimensional quantities $\{m_k(x) = \mathrm{E}(Y_t | X_{t,k} = x), 1 \leq k \leq L\}$ where $X_{t,k}$ is the k-th component of X_t and where L is a given number. The hypothesis could then be specified as

$$H_0' : \{m_k(\cdot), 1 \leq k \leq L\} \text{ is linear}$$

against

$$H_1' : \text{At least one } m_k(\cdot) \text{ is nonlinear.}$$

For Gaussian processes H_0 implies H_0', but there exist non-Gaussian ARMA processes for which this is not the case. Thus there are theoretical problems involved in comparing H_0 and H_0'. We will take a pragmatic view, and in practice reject the hypothesis of linearity if the difference between $m_k(\cdot)$ and the corresponding lag-k linear predictor is large. This corresponds roughly to looking at plots of the nonparametric regression at various lags and rejection means that there exists at least one k for which the lag-k nonlinear predictor is better than the lag-k linear predictor. As will be seen, the asymptotic theory is most easily derived in the case of H_0, however. Finally, as in Hjellvik and Tjøstheim (1995,1996), it should be noted that the bootstrap version of the test is constructed modulo an autoregressive or autoregressive moving average approximation in the first stage, so in this case it is H_0 which is tested based on functionals motivated by H_0'.

We construct the tests using the local polynomial regression estimator of $m_k(\cdot)$, and its derivatives. Locally at the point x, by a Taylor expansion of order T, we have

$$m_k(z) \approx \sum_{i=0}^{T} \frac{m_k^{(i)}(x)}{i!}(z-x)^i \tag{2.2}$$

where $m_k^{(i)}(x)$ denotes the i'th derivative of $m_k(x)$ (we will also use primes to denote the first and second derivative). Now consider the following least squares problem: Let $\hat{\gamma}_i,\ i = 0, \ldots, T$ minimize

$$\sum_{t=1}^{n} \left\{ Y_t - \sum_{i=0}^{T} \frac{\gamma_i}{i!}(X_{t,k}-x)^i \right\}^2 K\left(\frac{X_{t,k}-x}{h}\right), \tag{2.3}$$

where K is a nonnegative function, which serves as a kernel function, and h is the bandwidth, controlling the size of the local neighbourhood. Then, $\hat{\gamma}_i$ estimates $m_k^{(i)}(x),\ i = 0, \ldots, T$. Let $\gamma = (m_k(x), m_k^{(1)}(x), \ldots, m_k^{(T)}(x))^\tau$. The least square theory provides the solution

$$\hat{\gamma} = (X^\tau W X)^{-1} X^\tau W Y, \tag{2.4}$$

where $Y = (Y_1, \ldots, Y_n)^\tau$, $W = \mathrm{diag}(K_h(X_{1,k}-x), \ldots, K_h(X_{n,k}-x))$, $K_h(\cdot) = h^{-1}K(\cdot/h)$, and X is a $n \times (T+1)$ matrix with the i-th row $(1, (X_{i,k}-x), \ldots, (X_{i,k}-x)^T/T!)$. The special case with $T = 0$ corresponds to the ordinary kernel method of estimation. The theory of local polynomial regression has recently been developed in a number of papers (cf. Fan 1992 and 1993, Fan et al 1993, and Ruppert and Wand 1994).

If the model is linear in the sense that $m_k(x) = \theta_k x$, then $m_k'(x) \equiv \theta_k$ and $m_k''(x) \equiv 0$, and therefore we would expect

$$\hat{m}_k(x) \approx \hat{\theta}_k x, \quad \hat{m}_k'(x) \approx \hat{\theta}_k, \quad \hat{m}_k''(x) \approx 0, \qquad \text{for all } x \in R^d,$$

where $\hat{\theta}_k$ is the LSE of θ_k under H_0'. Based on this observation, we define the following statistics for testing the linearity of model (2.1):

$$\hat{L}_T(m_k) = \frac{1}{n}\sum_{t=1}^{n}(\hat{m}_k(X_{t,k}) - \hat{\theta}_k X_{t,k})^2 w(X_{t,k}), \quad T \geq 0 \tag{2.5}$$

$$\hat{L}_T(m_k') = \frac{1}{n}\sum_{t=1}^{n}(\hat{m}_k'(X_{t,k}) - \hat{\theta}_k)^2 w(X_{t,k}), \quad T \geq 1 \tag{2.6}$$

$$\hat{L}_T(m_k'') = \frac{1}{n}\sum_{t=1}^{n}\hat{m}_k''(X_{t,k})^2 w(X_{t,k}), \quad T \geq 2 \tag{2.7}$$

where $w(\cdot)$ is a continuous weight function. In fact, it can be proved that as the sample size tends to infinity, all of the above statistics converge to 0 when $m_k(\cdot)$ is linear (cf. Theorem 3.1 below).

Therefore large values of the statistics indicate possible departure from linear models. Following a suggestion by the referee, in principle the joint limit distribution of the statistics in (2.5)–(2.7) can be derived. This could be used to construct a suitable quadratic form involving the three statistics and use this as a single limiting chi-square statistic for linearity testing.

To ease the analytical derivations, we express the solution of (2.4) as follows:

$$\hat{m}_k^{(i)}(x) = \frac{i!}{nh^{(i+1)}} \sum_{t=1}^{n} W_{n,i}\left(\frac{X_{t,k}-x}{h}, x\right) Y_t, \ i = 0, \ldots, T \tag{2.8}$$

in which the vector function $W_{n,i}(\cdot, \cdot)$ is defined as

$$W_{n,i}(t,x) = e_{i+1}^{\tau} S_n^{-1}(x)(1, t, t^2, \ldots, t^T)^{\tau} K(t) \tag{2.9}$$

where e_i is the unit vector with the i'th element equal to 1, $S_n(x)$ is a $(T+1) \times (T+1)$ matrix with the (i,j)-th element s_{i+j-2}, and

$$s_j \equiv s_j(x) = \frac{1}{n} \sum_{t=1}^{n} \left(\frac{X_{t,k}-x}{h}\right)^j K_h(X_{t,k}-x). \tag{2.10}$$

3 Asymptotic Properties

There are at least two reasons for considering asymptotic properties. First, it is desirable to establish that our statistics have reasonable properties as $n \to \infty$, even though, as will be seen in Section 5, very large sample sizes are required to obtain a good approximation to the asymptotic distribution. The other reason for deriving asymptotics is the problem of choosing the bandwidth. Its connection with the asymptotic distribution is discussed in Section 4.

For reasons of simplicity and space we only consider the local quadratic regression ($T = 2$) for model (2.1) with a one-dimensional regressor ($d = 1$). The statistics of interest are the functionals defined as in (2.5)–(2.7) with $k = 1$. (Note that $m_1(\cdot) = m(\cdot)$ when $d = 1$.) For the cases with $d > 1$, the asymptotic results still hold but with more complicated notation.

We start by stating some regularity conditions:

(A1) The kernel function K is a symmetric density function with a bounded support in R, and $|K(x_1) - K(x_2)| \leq c|x_1 - x_2|$ for all x_1 and x_2 in its support. The weight function $w(\cdot)$ is continuous and with compact support contained in $\{p(x) > 0\}$.

(A2) For all t, $\mathrm{E}\{\, \epsilon_t \,|\, X_t, X_{t-1}, \ldots; Y_{t-1}, Y_{t-2}, \ldots\} = \mathrm{E}\{\, \epsilon_t \,|\, X_t \,\} = 0$. $\mathrm{E}\{X_t^8\} < \infty$, and $\mathrm{E}\,\{Y_t^8\} < \infty$. Further, $\mathrm{E}\{Y^2 | X = x\}$ is a bounded function of x.

(A3) The joint density of distinct elements of $(X_1, Y_1, X_s, Y_s, X_t, Y_t)$ $(t > s > 1)$ is continuous and bounded by a constant independent of s and t.

(A4) The process $\{(X_t, Y_t)\}$ is absolutely regular, i.e.

$$\beta(j) \equiv \sup_{i \geq 1} \mathrm{E} \left\{ \sup_{A \in \mathcal{F}_{i+j}^{\infty}} |P(A|\mathcal{F}_1^i) - P(A)| \right\} \to 0, \quad \text{as } j \to \infty,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $\{(X_k, Y_k) : k = i, \ldots, j\}$, $(j \geq i)$. Further, for a constant $\delta \in (0, 0.5)$, $\sum_{k=1}^{\infty} k^2 \beta^{\frac{\delta}{1+\delta}}(k) < \infty$.

(A5) As $n \to \infty$, then $h \to 0$ and $nh^{\frac{2+4\delta}{1+\delta}} / \log n \to \infty$.

An autoregressive process would satisfy (A2)–(A4) under mild assumptions on the generating mechanism. The condition on the boundedness of $K(\cdot)$ in (A1) is imposed for the brevity of proofs. The assumption on the convergence rates of h in (A5) is also for technical convenience. It can be weakened by applying Collomb's inequality (Lemma 2.2 of Györfi *et al.* 1989), which involves more technical details (cf. §2.3 of Györfi *et al.* 1989). The assumption of the convergence rates of $\beta(j)$ is also not the weakest possible. Note that under condition (A4), the process is strongly mixing (cf. Bradley 1986). Conditions (A2) and (A4) ensure that the bias of the estimators converge to zero (see Rios 1996).

Lemma 3.1. Under conditions (A1), (A4), and (A5), for $s_j(x)$ defined as in (2.10) $(j = 0, 1, \ldots, 4)$,

$$\sup_{x \in G} |s_j(x) - \mathrm{E}\{s_j(x)\}| \xrightarrow{P} 0$$

for any compact subset G in R.

Proof. We prove this only for the case with $j = 0$. For any $\varepsilon > 0$ and $x \in G$,

$$\begin{aligned} & P\{|s_0(x) - \mathrm{E}\{s_0(x)\}| \geq \varepsilon\} \leq \varepsilon^{-2} \mathrm{Var}\{s_0(x)\} \leq \frac{1}{n^2 \varepsilon^2} \sum_{i=1}^{n} \mathrm{E}\{K_h(X_i - x)\}^2 \\ + \quad & \frac{2}{nh^2\varepsilon^2} \sum_{j=1}^{n-1} \left[\mathrm{E}\left\{ K\left(\frac{X_1 - x}{h}\right) K\left(\frac{X_{j+1} - x}{h}\right) \right\} - \left\{ \mathrm{E} K\left(\frac{X_1 - x}{h}\right) \right\}^2 \right] \\ \leq \quad & c \left\{ \frac{1}{nh} + \frac{1}{nh^{2-c_0}} \sum_{j=1}^{n} \beta^{1-c_0/2}(j) \right\} \equiv \pi_n \end{aligned} \tag{3.1}$$

where $c_0 \in (0, 1)$ is a constant. The last inequality follows from Lemma 1 of Yoshihara (1976).

We cover G by a finite number of open intervals B_k centered at x_k in such a way that

$$G \subset \bigcup_{k=1}^{l_n} B_k, \qquad \sup_{x \in B_k} |x - x_k| \le h^2/\log n, \qquad l_n \le c_2 h^{-2} \log n, \tag{3.2}$$

where $c_2 > 0$ is a constant. Consequently, for $x \in B_k$, $|K_h(X_t - x) - K_h(X_t - x_k)| \le c/\log n$ for all X_t, where c is independent of k. Thus

$$\begin{aligned} & P\{\sup_{x \in G} |s_0(x) - \mathrm{E}\{s_0(x)\}| \ge \varepsilon\} \\ = \; & P\{\max_{1 \le k \le l_n} |s_0(x_k) - \mathrm{E}\{s_0(x_k)\}| + O((\log n)^{-1}) \ge \varepsilon\} \\ \le \; & l_n \pi_n + o(1). \end{aligned}$$

The last inequality follows from (3.1). Condition (A5) and (3.2) ensure that $l_n \pi_n \to 0$. The proof is completed.

A result similar to that of Lemma 3.1 has been established by Ango Nze and Rios (1995). It follows that for any compact $G \subset \{p(x) > 0\}$, uniformly for $x \in G$,

$$S_n^{-1}(x) \xrightarrow{P} p^{-1}(x) \begin{pmatrix} 1 & 0 & \mu_2 \\ 0 & \mu_2 & 0 \\ \mu_2 & 0 & \mu_4 \end{pmatrix}^{-1} = p^{-1}(x) \begin{pmatrix} \frac{\mu_4}{\mu_4 - \mu_2^2} & 0 & \frac{-\mu_2}{\mu_4 - \mu_2^2} \\ 0 & \frac{1}{\mu_2} & 0 \\ \frac{-\mu_2}{\mu_4 - \mu_2^2} & 0 & \frac{1}{\mu_4 - \mu_2^2} \end{pmatrix}, \tag{3.3}$$

where $\mu_j = \int t^j K(t) \mathrm{d}t$, $j \ge 1$.

¿From (2.4), we have that

$$\hat{\gamma} - \gamma = (X^\tau W X)^{-1} X^\tau W (Y - X\gamma).$$

Note that under H_0, $m(x) = \theta x$, $m'(x) \equiv \theta$, $m''(x) \equiv 0$, and $m(X_i) = m(x) + (X_i - x)m'(x)$. Similar to (2.8), under H_0, we have the expressions

$$\hat{m}(x) - \theta x = \frac{1}{nh} \sum_{i=1}^{n} W_{n,0}\left(\frac{X_i - x}{h}, x\right)$$

$$\hat{m}'(x) - \theta = \frac{1}{nh^2} \sum_{i=1}^{n} W_{n,1}\left(\frac{X_i - x}{h}, x\right) \epsilon_i, \quad \hat{m}''(x) = \frac{1}{nh^3} \sum_{i=1}^{n} W_{n,2}\left(\frac{X_i - x}{h}, x\right) \epsilon_i,$$

where ϵ_i is given as in (2.1). By (2.9) and (3.3) (cf. Rios 1996 for the bias calculations), uniformly for $x \in G$,

$$\hat{m}(x) - \theta x = \left\{ \frac{1}{nh(\mu_4 - \mu_2^2)p(x)} \sum_{i=1}^{n} \left\{ \mu_4 - \mu_2 \left(\frac{X_i - x}{h}\right)^2 \right\} \epsilon_i K\left(\frac{X_i - x}{h}\right) \right\} \{1 + o_p(1)\}, \tag{3.4}$$

$$\hat{m}'(x) - \theta = \left\{ \frac{1}{nh^2\mu_2 p(x)} \sum_{i=1}^{n} \left(\frac{X_i - x}{h} \right) \epsilon_i K \left(\frac{X_i - x}{h} \right) \right\} (1 + o_p(1)), \tag{3.5}$$

$$\hat{m}''(x) = \left\{ \frac{2}{nh^3(\mu_4 - \mu_2^2)p(x)} \sum_{i=1}^{n} \left\{ \left(\frac{X_i - x}{n} \right)^2 - \mu_2 \right\} \epsilon_i K \left(\frac{X_i - x}{h} \right) \right\} (1 + o_p(1)).$$

The next lemma plays a key role in deriving the asymptotical distribution of the statistics in testing nonlinearity and independence by using the local polynomial regression method. Hall (1984) and de Jong (1987) discussed similar results for independent observations.

To state our result, we introduce some notation. Suppose $\varphi_n(\cdot,\cdot)$ is a symmetric Borel function defined on $R^p \times R^p$, which may depend on the sample size n. We also assume that there exists a sequence of σ-algebras $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ for which $\xi_j \in \mathcal{F}_j$, and further

$$\text{(i)} \quad \mathrm{E}\{\varphi_n(x, \xi_1)\} = 0, \quad \text{for any } x \in R^p,$$

$$\text{(ii)} \quad \mathrm{E}\{\varphi_n(\xi_i, \xi_j)|\mathcal{F}_{j-1}\} = 0, \quad \text{for any } i < j.$$

The statistic of interest is in the form of

$$U_n = \sum_{1 \le i < j \le n} \varphi(\xi_i, \xi_j).$$

As pointed out by Hall (1984), U_n can be expressed as a partial sum of a sequence of martingale differences:

$$U_n = \sum_{k=2}^{n} V_k, \quad \text{and} \quad V_k = \sum_{i=1}^{k-1} \varphi_n(\xi_i, \xi_k). \tag{3.6}$$

The index n is suppressed in the notation V_k.

Let $\varphi_{ij} = \varphi_n(\xi_i, \xi_j)$, $\sigma_{ij}^2 = \mathrm{Var}(\varphi_{ij})$, and $\sigma_n^2 = \sum_{1 \le i < j \le n} \sigma_{ij}^2$. For some constant $\delta > 0$, define

$$M_{n1} = \max_{1<i<j\le n} \max \left\{ \mathrm{E}|\varphi_{1j}\varphi_{ij}|^{1+\delta}, \int |\varphi_{1j}\varphi_{ij}|^{1+\delta} \mathrm{d}P(\xi_1)\mathrm{d}P(\xi_i, \xi_j) \right\},$$

$$\begin{aligned} M_{n2} \quad = \quad & \max_{1<i<j\le n} \max \left\{ \mathrm{E}|\varphi_{1j}\varphi_{ij}|^{2(1+\delta)}, \int |\varphi_{1j}\varphi_{ij}|^{2(1+\delta)} \mathrm{d}P(\xi_1)\mathrm{d}P(\xi_i, \xi_j), \right. \\ & \left. \int |\varphi_{1j}\varphi_{ij}|^{2(1+\delta)} \mathrm{d}P(\xi_1, \xi_i)\mathrm{d}P(\xi_j), \int |\varphi_{1j}\varphi_{ij}|^{2(1+\delta)} \mathrm{d}P(\xi_1)\mathrm{d}P(\xi_i)\mathrm{d}P(\xi_j) \right\}, \end{aligned}$$

$$M_{n3} = \max_{1 \le i < j \le n} \mathrm{E}|\varphi_{1j}\varphi_{ij}|^2, \qquad M_{n4} = \max_{\substack{1<i,j,k\le 2n \\ i,j,k \text{ different}}} \left\{ \max_P \int |\varphi_{1i}\varphi_{jk}|^{2(1+\delta)} \mathrm{d}P \right\}, \tag{3.7}$$

$$M_{n5} = \max_{j>i>1} \max \left\{ \mathrm{E} \left| \int \varphi_{1i}\varphi_{1j} \mathrm{d}P(\xi_1) \right|^{2(1+\delta)}, \int \left| \int \varphi_{1i}\varphi_{1j} \mathrm{d}P(\xi_1) \right|^{2(1+\delta)} \mathrm{d}P(\xi_i)\mathrm{d}P(\xi_j) \right\},$$

$$M_{n6} = \max_{j>i>1} \mathrm{E} \left| \int \varphi_{1i}\varphi_{1j} \mathrm{d}P(\xi_1) \right|^2,$$

where the maximization over P in the equation for M_{n4} is taken over the four probability measures $P(\xi_1, \xi_i, \xi_j, \xi_k)$, $P(\xi_1)P(\xi_i, \xi_j, \xi_k)$, $P(\xi_1)P(\xi_{i_1})P(\xi_{i_2}, \xi_{i_3})$, and $P(\xi_1)P(\xi_i)P(\xi_j)P(\xi_k)$, where (i_1, i_2, i_3) is the permutation of (i, j, k) in the ascending order. We assume that all of the above constants are finite.

Lemma 3.2. If for some $\delta > 0$, $\sum_{k=1}^{\infty} k^2\{\beta(k)\}^{\frac{\delta}{1+\delta}} < \infty$, and

$$\max \frac{1}{\sigma_n^2}\left\{n^2\{M_{n1}^{\frac{1}{1+\delta}} + M_{n5}^{\frac{1}{2(1+\delta)}} + M_{n6}^{\frac{1}{2}}\},\ n^{\frac{3}{2}}\{M_{n2}^{\frac{1}{2(1+\delta)}} + M_{n3}^{\frac{1}{2}} + M_{n4}^{\frac{1}{2(1+\delta)}}\}\right\} \longrightarrow 0, \tag{3.8}$$

as $n \to \infty$, where $M_{n1}, \ldots, M_{n4}$ are defined as in (3.7), then $\sigma_n^{-1}U_n$ is asymptotically normal with mean value 0 and variance 1.

The proof of this central limit result is lengthy and it is contained in Appendix 1 of Hjellvik *et al.* (1996).

The following three theorems present the asymptotic behavior of the statistics $\hat{L}_2(m)$, $\hat{L}_2(m')$ and $\hat{L}_2(m'')$, defined in (2.5)–(2.7). We use the notation $X \sim AN(\mu, \sigma^2)$ to denote that $\sigma^{-1}(X - \mu) \xrightarrow{d} N(0, 1)$.

Theorem 3.1. Let conditions (A1)–(A5) hold. Under the hypothesis H_0, $\hat{L}_2(m) \xrightarrow{P} 0$, $\hat{L}_2(m') \xrightarrow{P} 0$ and $\hat{L}_2(m'') \xrightarrow{P} 0$ as $n \to \infty$.

Theorem 3.2. Let conditions (A1)–(A5) hold. Under the hypothesis H_0,

$$\text{(i)}\ \hat{L}_2(m) \sim AN(a_1/(nh), \sigma_1^2/(n^2h)), \tag{3.9}$$

$$\text{(ii)}\ \hat{L}_2(m') \sim AN(a_2/(nh^3), \sigma_2^2/(n^2h^5)), \tag{3.10}$$

$$\text{(iii)}\ \hat{L}_2(m'') \sim AN(a_3/(nh^5), \sigma_3^2/(n^2h^9)), \tag{3.11}$$

where

$$\begin{aligned} a_1 &= \frac{1}{(\mu_4 - \mu_2^2)^2}\int(\mu_4 - \mu_2 t^2)^2 K^2(t)\mathrm{d}t \int \sigma^2(x)w(x)\mathrm{d}x, \\ a_2 &= \frac{1}{\mu_2^2}\int t^2 K^2(t)\mathrm{d}t \int \sigma^2(x)w(x)\mathrm{d}x, \\ a_3 &= \frac{4}{(\mu_4 - \mu_2^2)^2}\int(t^2 - \mu_2)^2 K^2(t)\mathrm{d}t \int \sigma^2(x)w(x)\mathrm{d}x, \end{aligned}$$

$$\sigma_1^2 = \frac{2}{(\mu_4 - \mu_2^2)^4}\int \sigma^4(x)w^2(x)\mathrm{d}x \int(\mu_4 - \mu_2 u^2)(\mu_4 - \mu_2 v^2)\{\mu_4 - \mu_2(u - z)^2\}$$

$$\times\{\mu_4 - \mu_2(v-z)^2\}K(u)K(v)K(u-z)K(v-z)\mathrm{d}u\mathrm{d}v\mathrm{d}z,$$

$$\sigma_2^2 = \frac{2}{\mu_2^4}\int \sigma^4(x)w^2(x)\mathrm{d}x \int uv(u-z)(v-z)K(u)K(v)K(u-z)K(v-z)\mathrm{d}u\mathrm{d}v\mathrm{d}z,$$

$$\sigma_3^2 = \frac{32}{(\mu_4-\mu_2^2)^4}\int \sigma^4(x)w^2(x)\mathrm{d}x \int (u^2-\mu_2)(v^2-\mu_2)\{(u-z)^2-\mu_2\}$$
$$\times\{(v-z)^2-\mu_2\}K(u)K(v)K(u-z)K(v-z)\mathrm{d}u\mathrm{d}v\mathrm{d}z.$$

In the above expressions, $\mu_j = \int t^j K(t)\mathrm{d}t$, and $\sigma^2(x) = \mathrm{Var}(Y_1|X_1 = x)$.

Theorem 3.3. Let conditions (A1)–(A5) hold. Under the hypothesis H_1,

$$\text{(i)}\ \hat{L}_2(m) \sim AN(a_4, \sigma_4^2/n),$$
$$\text{(ii)}\ \hat{L}_2(m') \sim AN(a_5, \sigma_5^2/n),$$
$$\text{(iii)}\ \hat{L}_2(m'') \sim AN(a_6, \sigma_6^2/n),$$

where

$$a_4 = \mathrm{E}\{[m(X_1) - \theta X_1]^2 w(X_1)\} + o(1), \quad \text{and} \quad \sigma_4^2 = \mathrm{Var}\{[m(X_1) - \theta X_1]^2 w(X_1)\},$$
$$a_5 = \mathrm{E}\{[m'(X_1) - \theta]^2 w(X_1)\} + o(1), \quad \text{and} \quad \sigma_5^2 = \mathrm{Var}\{[m'(X_1) - \theta]^2 w(X_1)\},$$
$$a_6 = \mathrm{E}\{m''(X_1)^2 w(X_1)\} + o(1), \quad \text{and} \quad \sigma_6^2 = \mathrm{Var}\{m''(X_1)^2 w(X_1)\}.$$

In the above expressions, $\theta = \{\mathrm{Var}(X_1)\}^{-1}\mathrm{Cov}(X_1, Y_1)$.

Theorem 3.1 shows that all of the test statistics will converge to zero in probability if the model is linear. Therefore, large values of the statistics will indicate departure from the linearity hypothesis. The asymptotic approximations for the significance level and power of the tests $\hat{L}_2(m)$, $\hat{L}_2(m')$ and $\hat{L}_2(m'')$ are given in Theorems 3.2 and 3.3, respectively. The main idea in the proofs of all of the theorems is to perform the Hoeffding's decomposition on the test statistics and then apply asymptotic results for U-statistics (cf. Yoshihara 1976, Denker and Keller 1983, and Lemma 3.2 above. For example, under the hypothesis H_0, the statistic $\hat{L}_2(m')$ is asymptotically equivalent to the sum of a constant and a quadratic form (see (a) – (e) in Appendix 2 of Hjellvik *et al.* 1996). Theorem 3.2 shows that $nh^{\frac{5}{2}}\hat{L}_2(m')$ is approximately normal with variance σ_2^2 , but the mean diverges to infinity. The asymptotics in such a form have been noted before in similar problems (cf. Proposition 1 of Härdle and Mammen 1993). The results of Theorem 3.3 are somehow trivial. One could instead assume that the alternative hypothesis H_1 deviates from H_0 at a certain rate related to the sample size (cf. Härdle and Mammen 1993). Since it is arguable whether such a

hypothesis has any practical implications in the current context, we do not pursue this any further. A proof of Theorem 3.2 (ii) is given in Appendix 2 of Hjellvik *et al.* (1996).

4 The Role Played by the Bandwidth

It has been pointed out that local polynomial regression methods perform better in various respects than some conventional kernel regression methods (cf. Fan 1992 and 1993, Hastie and Loader 1993). In this section, we point out two interesting features in using the local quadratic regression method in linearity testing.

To see how the bandwidth h varies in local polynomial regressions with the different orders of the polynomials, let us consider the estimators of $m(\cdot)$ in the case that $d = 1$. For the local quadratic regression, similar to (3.4), for $x \in \{p(x) > 0\}$, it can be proved that

$$\hat{m}(x)-m(x) = \left\{ \frac{1}{nh(\mu_4 - \mu_2^2)p(x)} \sum_{i=1}^{n} \left\{ \mu_4 - \mu_2 \left(\frac{X_i - x}{h} \right)^2 \right\} \{\epsilon_i + \eta_i(x)\} K \left(\frac{X_i - x}{h} \right) \right\} \{1+o_p(1)\},$$

where

$$\eta_i(x) = m(X_i) - m(x) - m'(x)(X_i - x) - \frac{1}{2} m''(x)(X_i - x)^2.$$

By Theorem 7.8.4 of Shiryayev (1984), it can be shown that

$$\sqrt{nh} \left\{ \hat{m}(x) - m(x) - h^4 \frac{\mu_4^2 - \mu_2\mu_4}{24(\mu_4 - \mu_2^2)} \{m^{(4)}(x) + 2m^{(3)}(x)p'(x)/p(x)\} \right\}$$

$$\xrightarrow{d} N \left(0, \frac{\sigma^2(x)(\mu_4 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4)}{p(x)(\mu_4 - \mu_2^2)^2} \right),$$

where $m^{(k)}$ denotes the k-th derivative of $m(\cdot)$, and $\nu_k = \int t^k K^2(t) \mathrm{d}t$ (cf. Appendix 2 of Hjellvik *et al.* 1996). Hence, the approximate MSE of the estimator can be defined as

$$\begin{aligned} \mathrm{MSE}_2(h) &= h^8 \left(\frac{\mu_4^2 - \mu_2\mu_4}{24(\mu_4 - \mu_2^2)} \right)^2 \{m^{(4)}(x) + 2m^{(3)}(x)p'(x)/p(x)\}^2 \\ &\quad + \frac{1}{nh} \frac{\sigma^2(x)(\mu_4 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4)}{p(x)(\mu_4 - \mu_2^2)^2}. \end{aligned}$$

Hence, the (theoretical) optimal bandwidth which minimizes the above $\mathrm{MSE}_2(h)$ is of the order

$$h_2(x) \propto \frac{1}{n^{1/9}}. \tag{4.1}$$

If we impose the constraint $\gamma_i = 0, i \geq 2$ in the minimization of (2.3), the estimator derived is a local linear regression estimator of $m(\cdot)$. Further, if in addition γ_1 is restricted to 0, we

get the local constant regression estimator of $m(\cdot)$, which is the conventional Nadaraya-Watson kernel estimator. We will see that for local linear or local constant regression estimators, the optimal bandwidths should be smaller than $h_2(x)$. In fact, based on the asymptotic normality, the approximate MSE of the local linear regression estimator of $m(\cdot)$ is

$$\mathrm{MSE}_1(h) = h^4 \frac{\mu_2^2}{4}\{m''(x)\}^2 + \frac{1}{nh}\frac{\sigma^2(x)\nu_0}{p(x)}.$$

To minimize $\mathrm{MSE}_1(h)$, we should use the bandwidth

$$h_1(x) = \frac{1}{n^{1/5}}\left(\frac{\sigma^2(x)\nu_0}{p(x)\mu_2^2\{m''(x)\}^2}\right)^{1/5}. \tag{4.2}$$

Similarly, if we use the Nadaraya-Watson kernel estimator of $m(\cdot)$. The approximate MSE is

$$\mathrm{MSE}_0(h) = h^4 \frac{\mu_2^2}{4}\left\{m''(x) + \frac{4m'(x)p'(x)}{p(x)}\right\}^2 + \frac{1}{nh}\frac{\sigma^2(x)\nu_0}{p(x)}$$

(cf. also Ango Nze and Doukhan 1993). The optimal bandwidth which minimizes $\mathrm{MSE}_0(h)$ is

$$h_0(x) = \frac{1}{n^{1/5}}\left(\frac{\sigma^2(x)\nu_0}{p(x)\mu_2^2\{m''(x) + 4m'(x)p'(x)/p(x)\}^2}\right)^{1/5}. \tag{4.3}$$

It is seen that $h_i(x)/h_2(x) \to 0$ as $n \to \infty$ for $i = 0, 1$. Further, although $h_0(x)$ and $h_1(x)$ are of the same order, $h_0(x)$ is usually smaller than $h_1(x)$. Note that one more term $4m'(x)p'(x)/p(x)$ appears in the denominator of the right hand side of (4.3), which is due to the larger bias of the Nadaraya-Watson estimator compared with the local linear regression estimator (see also Fan 1992). The above observation indicates that the bandwidth should be increased when the order of the polynomial in a local polynomial regression estimation is increased. Intuitively, it is easy to understand. For example, locally around x, a quadratic function can fit the curve $m(\cdot)$ in a larger neighbourhood of x 'as well as' a linear function could fit $m(\cdot)$ in a smaller neighbourhood. Therefore, local quadratic fit can encompass more local variation in the data.

Another advantage of using the local quadratic regression method in our tests is that they become less sensitive to the choice of bandwidth, in particular they are more robust to oversmoothing.

To see the effect of a very large bandwidth, consider the following extreme case. For a given sample $\{(X_t, Y_t),\ 1 \le t \le n\}$, we minimize (2.3) with $T = 2$ to estimate the functions of interest. Suppose we let h tend to infinity, and that the kernel function K is chosen so that $K_h(\cdot)$ converges to a positive constant as $h \to \infty$. Then (2.3), with $T = 2$, is proportional to

$$\sum_{t=1}^{n}\{Y_t - a - b(X_{t,k} - x) - c(X_{t,k} - x)^2\}^2$$

$$= \sum_{t=1}^{n} \{Y_t - \alpha_1 - \alpha_2 X_{t,k} - \alpha_3 X_{t,k}^2\}^2,$$

where α_1, α_2 and α_3 are properly defined constants. Now, to minimize (2.3) over (a, b, c) is equivalent to the problem of minimizing the above function over $(\alpha_1, \alpha_2, \alpha_3)$, which means that the local quadratic regression reduces to regress Y_t as a global quadratic function of $X_{t,k}$. If the function $m_k(\cdot)$ is not linear, this global parametric fitting could be capable of showing the departure from the linearity hypothesis. This explains why our test statistics with large values of h still have power to detect the nonlinearity, which is illustriously different from the statistics based on local linear or local constant regression estimators. Cox (1981) suggested the use of quadratic or cubic regression to test nonlinearity.

The larger robustness of local polynomial based tests to the choice of bandwidth is confirmed by simulation experiments reported in Section 6.

5 Evaluation of finite sample properties of the tests

Based on the experience of Hjellvik and Tjøstheim (1995,1996), it is essential to examine the finite sample properties of the tests. We have done a number of simulation experiments in a time series setting, and it is convenient to introduce the notation $M_k(x) = \mathrm{E}\{X_t | X_{t-k} = x\}$, which was used in Hjellvik and Tjøstheim (1995,1996), as a special case of $m_k(x) = \mathrm{E}\{Y_t | X_{t,k} = x\}$. The corresponding sum of squares to be minimized for a T-th order local polynomial approximation is

$$\sum_{t=k+1}^{n} \left\{ X_t - \sum_{i=0}^{T} \frac{\gamma_i}{i!} (X_{t-k} - x)^i \right\}^2 K\left(\frac{X_{t-k} - x}{h}\right).$$

We have used a Gaussian kernel function here and elsewhere in the paper. To detect nonlinearity in the conditional variance, we also introduce $V_k(x) = \mathrm{var}(X_t | X_{t-k} = x)$. Linearity in terms of the conditional variance is taken to mean that $V_k(x) \equiv c$ where c is a positive constant, and $V_k'(x) \equiv 0$. However, as in Hjellvik and Tjøstheim (1995,1996) we conduct the variance tests on the residual process $\{e_t\}$ from the best autoregressive fit, rather than on $\{X_t\}$ itself. We then get $V_k(e) = \mathrm{var}(e_t | e_{t-k} = e)$ where $e_t = X_t - a_1 X_{t-1} - \ldots - a_p X_{t-p}$ for an autoregressive approximation of order p. Note that $\{e_t\}$ denotes the residual process from the autoregressive approximation whereas $\{\epsilon_t\}$ denotes the true residual process, and $V_k(e)$ may be non-constant even though $V_k(\epsilon)$ is constant. When using the test functionals, we standardize both $\{X_t\}$ and

$\{e_t\}$ so that they have zero mean and variance one. We now have the following statistics:

$$\hat{L}_T(M_k) = \frac{1}{n}\sum_{t=1}^{n}(\hat{M}_k(X_t) - \hat{\rho}_k X_t)^2 w(X_t), \quad T \geq 0 \tag{5.1}$$

$$\hat{L}_T(M_k') = \frac{1}{n}\sum_{t=1}^{n}(\hat{M}_k'(X_t) - \hat{\rho}_k)^2 w(X_t), \quad T \geq 1 \tag{5.2}$$

$$\hat{L}_T(M_k'') = \frac{1}{n}\sum_{t=1}^{n}\hat{M}_k''(X_t)^2 w(X_t), \quad T \geq 2 \tag{5.3}$$

$$\hat{L}_T(V_k) = \frac{1}{n-\hat{p}}\sum_{t=\hat{p}+1}^{n}(\hat{V}_k(\hat{e}_t) - \sigma_{\hat{e}}^2)^2 w(\hat{e}_t), \quad T \geq 0 \tag{5.4}$$

$$\hat{L}_T(V_k') = \frac{1}{n-\hat{p}}\sum_{t=\hat{p}+1}^{n}\hat{V}_k'(\hat{e}_t)^2 w(\hat{e}_t), \quad T \geq 1 \tag{5.5}$$

where $\hat{p}$ is the estimated order of the best autoregressive fit, ρ_k is the autocorrelation between X_t and X_{t-k} and $\sigma_{\hat{e}}^2$ is the empirical variance of $\{e_t\}$, which is equal to 1 due to the standardization. As an upper limit for $\hat{p}$ we have used $n/10$, and as weight function we have used the trapezoidal function $w(x) = 1(|x| \leq k) + (3 - 2|x|/k)\ 1(k < |x| \leq 3k/2)$ with $k = 2$ in the bootstrap approach and $k = 1$ in the evaluation of the asymptotic theory. Other weight functions have also been tried with roughly similar results.

We only examine the difference between the asymptotic distribution and the finite sample distribution in the trivial case of a Gaussian white noise process $\{X_t\} = \{\epsilon_t\}$, where $\{\epsilon_t\}$ consists of *iid* random variables. Such a simple example suffices to demonstrate the poorness of the asymptotic approximation. The results are given in Table 1, and it is seen that even for very large sample sizes the approximation is not very good. In fact a rectangular weight function on $[-1, 1]$ gave somewhat better results, the empirical sizes for $\hat{L}_2(M_1)$, $\hat{L}_2(M_1')$ and $\hat{L}_2(M_1'')$ being 0.032, 0.062 and 0.066, respectively for a sample size of $n = 2^{21}$. Note that the tabulated empirical sizes give a direct indication of the convergence in distribution implied by Theorem 3.2. We believe that the slow convergence is typical, as indicated by experiments in Hjellvik and Tjøstheim (1995,1996). We think that the reason for the bad approximation is that unlike a standard parametric situation, the next order terms in the Edgeworth expansion of our statistics are very close to the leading normal approximation terms given in Theorem 3.2. An example illustrating this phenomenon for a similar nonparametric test functional is given in Skaug and Tjøstheim (1993). Better approximation in a spesial case using a fixed experimental design have been reported by Poggi and Portier (1996).

6 The bootstrap approach

The outcome of the experiment in the preceding section means that for small and moderate sample sizes the asymptotic distribution cannot be used to construct the null-distribution of the functionals. An alternative is to create the null-distribution and the critical region corresponding to a nominal significance level by bootstrapping the residuals

$$\hat{e}_t = X_t - \sum_{i=1}^{\hat{p}} \hat{a}_i X_{t-i}$$

from the best linear autoregressive (or ARMA) fit to $\{X_t\}$. Bootstrapped values $\hat{L}^*(\cdot)$ of the functional in the null-situation are created by inserting bootstrapped linear versions

$$X_t^* = \sum_{i=1}^{\hat{p}} \hat{a}_i X_{t-i}^* + e_t^*$$

of $\{X_t\}$. By taking a sufficiently large number of bootstrap replicas $\{e_t^*\}$ of $\{\hat{e}_t\}$, in this way we can construct a null-distribution for $\hat{L}^*(\cdot)$. Both the conditional mean and the conditional variance functionals can be treated in this way, and for more details we refer to Hjellvik and Tjøstheim (1995,1996).

In the following, as a standard, 40 bootstrap replicas will be used to create the null-distribution. This distribution is smoothed by a nonparametric integrated kernel type estimate (cf. Hjellvik and Tjøstheim 1996), and the critical region corresponding to an α-level test thus is obtained by selecting the $(1-\alpha)$-quantile c_α^* of the bootstrap distribution, and the hypothesis of linearity is rejected if $\hat{L}(\cdot) \geq c_\alpha^*$, where $\hat{L}(\cdot)$ is the value of $\hat{L}$ as computed from the original data series $\{X_t\}$.

Ideally one would like to evaluate the merits of the bootstrap approach theoretically. Asymptotic theory is lacking for the bootstrap in this situation, although it should be possible to develop such a theory (Franke, Kreiss, Mammen 1995). However, this is a highly non-trivial task, which may require large sample sizes to be accurate. Possibly one should rather look at the randomization test aspect of the bootstrap and consider the approximation obtained using bootstrapped estimated residuals instead of permuted true residuals. If the latter were known, exact p-values could be obtained (cf. Skaug and Tjøstheim 1996). In the absence of a suitable theory we have reverted to an examination of the bootstrap approach by an extensive set of simulation experiments and by application to a wide range of real data sets. Much of the emphasis will be on the choice of bandwidth and the new aspects brought in by using local polynomial approximation. A power experiment on a wide class of nonlinear models listed in Table 2 has been conducted in Section 6.3.

Plots of the conditional mean and variance for these models are depicted in Figures 1 and 2 of Hjellvik *et al.* (1996).

6.1 The role played by the bandwidth

In Hjellvik and Tjøstheim (1995,1996) as a standard we used the bandwidth $h = n^{-1/5}$ on a normalized data series having zero mean and variance one. The results of Section 4 suggest that a more flexible approach to choosing the bandwidth should be taken in the case of local polynomial approximation. Figure 1 depicts the empirical size of the tests as a function of bandwidth for the statistics given in (5.1)–(5.3) for $T = 0, \ldots, 4$. The nominal size is 0.05, and the bootstrap is used to form the null-distribution as described above. The model used is linear autoregressive

$$X_t = aX_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0,1) \tag{6.1}$$

with $a = 0.5$. By comparison to Table 1 it is seen that the results obtained represent a vast improvement over those which could be achieved using asymptotic theory. In a sense this could be expected when it is taken into consideration that the bootstrap approach comes close to being a randomization test. It is seen that $\hat{L}_0(M_1)$ collapses when $h \geq 1.0$, whereas the other statistics seem to be quite independent of h. This can be expected since $T = 0$ corresponds to using the kernel method, and for h large, everything will be smoothened flat, i.e. $\hat{M}(x) \approx 0$, whereas $\hat{\rho}x \neq 0$, and the procedure breaks down. As indicated in Section 4, for linear models this does not happen for $T \geq 1$, as we then get global parametric estimates as $h \to \infty$. In practice $h = \infty$ is achieved by setting the kernel function $K(\cdot) \equiv 1$ in (2.3).

The empirical power clearly depends much more on h, as can be seen from Figure 2. The model used here is exponential autoregressive

$$X_t = (0.5 + b\exp(-0.5X_{t-1}^2))X_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0,1) \tag{6.2}$$

with $b = 1.3$ and we see as a general trend that the optimal value of h increases as the order T in the Taylor expansion increases. This is consistent with the results of Section 4 although those results were derived with another optimality criterion. We also see that for a given T, the optimal h increases with the derivatives. This is in accordance with general nonparametric estimation theory as a derivative of a regression estimate typically should be smoothed more (cf. Fan et al 1993). For $\hat{L}_0(M_1)$, the power drops quite fast for the same reason as for the level.

For h very large, $\hat{L}_0(M_1)$ and $\hat{L}_1(M_1)$ have no power since the parametric approximation to $M_1(x)$ obtained as $h \to \infty$ in Section 4 is constant and linear, respectively. Due to the specific

symmetry properties of model (6.2), also the quadratic approximation turns out to be nearly linear, so the power for $\hat{L}_2(M_1)$ is rather low. But for $T = 3$ and $T = 4$ the parametric approximation gives good results. For model lc) of Table 2, however, where $M_1(x)$ is approximately quadratic (see Figure 1 of Hjellvik *et al.* 1996), as can be expected the best result is achieved with $T = 2$ and $h = \infty$.

For the $\hat{L}(V_1)$-tests the size tends to be too low. Similar results were obtained in Hjellvik and Tjøstheim (1995,1996). Still the results are much better than those obtained using asymptotic theory.

Figure 3 shows the power of the $\hat{L}(V)$-tests for model la) of Table 2, and we see the same general trend as for the $\hat{L}(M)$-tests; the optimal h increases with T and the derivative. Here $\hat{L}_1(V_1)$ also has some power for $h = \infty$ because $V_1(\cdot)$ is constant under the null hypothesis. The estimated functional $\hat{L}_0(V_1)$ is much more robust than $\hat{L}_0(M_1)$, and this is the case for the other models listed in Table 2 as well.

6.2 Cross-validation of the bandwidth

When we cross-validate the bandwidth, we choose the value of h that gives the least prediction error. That is, we choose the h that minimizes

$$\sum_{k=1}^{K} \sum_{t=k+1}^{n} \{X_t - \hat{M}_k^-(X_{t-k})\}^2 w(X_{t-k}) \tag{6.3}$$

where $\hat{M}_k^-(X_{t-k})$ is the leave-k-out kernel estimator based on the data $(X_1, X_{k+1}), \ldots, (X_{t-k-1}, X_{t-1})$, $(X_{t+1}, X_{t+k+1}), \ldots, (X_{n-k}, X_n)$. We have used $K = 3$ in the following.

Taking $h = \infty$ means that we get a parametric approximation to $M(x)$: a polynomial of order T. So it is of interest to see whether $h = \infty$ is chosen by the cross-validation procedure when we have a linear model and use $T = 1$ or when we have a white noise model and use $T = 0$. (Of course $h = \infty$ is precluded from the asymptotic theory of Sections 3 and 4. Moreover, a polynomial $M(\cdot)$ with $T > 1$ would result in a transient time series model). Figure 6 in Hjellvik *et al.* (1996) shows that for model (6.1) $h = \infty$ is chosen in approximately 65 % of the cases relatively independent of the value of a when we use $T = 1$. Increasing n to 500 gives nearly the same result. For a negative coeffisient a, the result is approximately as for a positive. With $T = 0$, the percentage of $h = \infty$ is 63.4, 27.4 and 2.0 for $a = 0.0$, 0.2 and 0.4, respectively. The decreasing percentage as the coefficient a increases in value again is expected since for the kernel method $\hat{M}(x) \to 0$ as $h \to \infty$.

The bootstrap test is not well-defined for $T = 0$ or $T = 1$ combined with $h = \infty$. In particular when we have a nonlinear model, we do of course not want $h = \infty$ to be chosen when $T = 0$ or $T = 1$, but with a small autocorrelation, this may well happen for $T = 0$. In fact $h = \infty$ was chosen in 136 of 500 realizations of model lc) of Table 2 which is clearly nonlinear (cf. Figure 1 of Hjellvik *et al.* 1996). Similarly, $h = \infty$ was chosen in many cases with relatively strong nonlinearity of model (6.2). This shows a weakness of the cross validation procedure if h is allowed to vary freely. Traditionally a much more restricted range of h-values is used, and there are both logical and empirical reasons for not including $h = \infty$ in the allowable set of h's for $T \leq 1$. The routine implemented in the test, chooses h among the values in the upper part of Table 3. We do not cross-validate h for each bootstrap replica, but use the same h as for the original dataset.

For the $\hat{L}(V)$-tests we choose the h that minimizes

$$\sum_{k=1}^{K} \sum_{t=k+1}^{n} \{e_t^2 - \hat{M}_k^-(e_{t-k}^2)\}^2 w(e_{t-k}) \tag{6.4}$$

Here $h = \infty$ may give som power for $T = 1$ since the conditional variance is constant, under H_0, and we choose h according to the last two lines of Table 3.

Cross-validating with these values of h generally produced sizes somewhat higher than with h data independent, but more experiments are needed to make this firmer.

6.3 A power experiment for a wide set of models

We have performed a power experiment for the models listed in Table 2, where $\epsilon_t \sim N(0, 0.6^2)$ in model ld) - lf), $\epsilon_t \sim N(0, 0.7^2)$ in lg) - lj) and $\epsilon_t \sim N(0, 1)$ in the other models. Models la) - lj), aa) - ag) and Aa) - Ag) are discussed in Luukkonen *et al.* (1988), Ashley *et al.* (1986) and An and Cheng (1991), respectively. In this paper we mainly restrict ourselves to look at lag 1. For some of the models the nonlinearity is revealed at higher order lags.

Figure 4 shows the power for $\hat{L}_T(M_1)$ with h cross-validated for the models listed in Table 2 with $T = 0, \ldots, 3$. The + symbol indicates the power achieved in Hjellvik and Tjøstheim (1995). In most cases this is higher than the cross-validated power for $\hat{L}_0(M_1)$. One explanation is that the cross-validation procedure tends to pick out too large h's. Actually, in a rerun of the experiment of Figure 4 with a fixed bandwidth, on average somewhat better power was obtained.

Figure 5 shows the power for $\hat{L}_T(V_1)$ with h cross-validated for $T = 0, \ldots, 3$. Models lb) - lj) more or less have a constant conditional variance, and we have no comparable results in Hjellvik and Tjøstheim (1995). Therefore we have not run the $\hat{L}(V)$ tests on these. For the other

models $T = 2$ seems on average to be the best choice. Globally this corresponds to an ARCH-type dependence, which is seen from Figure 2 in Hjellvik *et al.* (1996) to be the predominant behavior of the conditional variance function. But we also see that $\hat{L}_0(V_1)$ in most cases gives a *higher* power than that achieved in Hjellvik and Tjøstheim (1995). This is probably because the bandwidth $h = n^{-1/5}$ used there is to small for the variance test. For some models (ae,Ab,Af) there is a dramatic improvement in power compared to Hjellvik and Tjøstheim (1995). Using derivatives on average gave inferior results.

6.4 The size for low nominal levels

So far we have far used a nominal level of 0.05 where our procedure of using 40 bootstrap replicas and the nonparametric approximation to the null distribution described by equation (3.5) in Hjellvik and Tjøstheim (1996) works quite satisfactory. At low nominal levels (≤ 0.01) however, this method gives a too high empirical level since the nonparametric estimate applied behaves rather poorly in the tails of the null distribution. This means that low p-values can not be trusted, and for a given real data set it will be difficult to interpret what a very low p-value really means. There is at least two ways in which this problem can be countered. The most obvious is to use more bootstrap replicas, but even at a nominal level of 0.001 more than 500 bootstrap replicas are needed, and it increases computer time. An effective way to reduce the number of bootstrap replicas is to apply the adaptive nonparametric density estimate described in Silverman (1985, p. 100 ff). This procedure uses a broader kernel in regions of low density, so the estimate of the tails of the null distribution becomes smoother, and we get some probability mass squeezed into the area in which we are interested. As can be seen from Figure 6 we now need only about 100 bootstrap replicas to get reasonable results for a nominal level of 0.001 for the white noise example. With 500 bootstrap replicas we can even approximate a nominal level of 0.00001. In fact using 500 bootstrap replicas leads to a too *low* empirical level at nominal levels 0.01, 0.001 and 0.0001. Clearly, there are many theoretical and practical problems left to be solved, and a separate investigation is really needed.

7 Some real data sets

In Hjellvik and Tjøstheim (1995) we presented some p-values for three real datasets: the sunspot data ($n = 361$), the lynx data ($n = 114$) and the blowfly data ($n = 288$). Some of these p-values were very low. Considering the results in Section 6.4 they were probably *too* low since they were

based on only 80 bootstrap replicas and the non-adaptive approximation to the null distribution. In Hjellvik and Tjøstheim (1995) we considered more general test statistics, which in the notation of this paper are defined by

$$\hat{L}_{T,\sup}(M_{10}) = \sup_{k\leq 10} \hat{L}_T(M_k), \qquad \hat{L}_{T,\text{ave}}(M_{10}) = \frac{1}{10}\sum_{k=1}^{10} \hat{L}_T(M_k),$$

$$\hat{L}_{T,\sup}(V_{10}) = \sup_{k\leq 10} \hat{L}_T(V_k), \qquad \hat{L}_{T,\text{ave}}(V_{10}) = \frac{1}{10}\sum_{k=1}^{10} \hat{L}_T(V_k).$$

If we compare the results in Table 4 based on 500 bootstrap replicas and the adaptive approximation to those in Table 8 in Hjellvik and Tjøstheim (1995), we see that for all but one of the cases in Table 8 with p-values less than 0.01, the corresponding p-values in Table 4 are higher. In Figure 7 the bandwidth is cross-validated, and the results for $T \geq 0$ included. For the log(lynx) data $\hat{L}_{0,\sup}(V_{10})$ and $\hat{L}_{0,\text{ave}}(V_{10})$ give rejection of linearity at level 0.05 in Figure 7, but not in Table 4. This is because the estimated order of the best autoregressive fit in Figure 7 is 11 ($= n/10$), whereas the upper limit in Table 4 is 10 (= the maximum order of Table 8 in Hjellvik and Tjøstheim 1995) and the estimated order is 2.

Acknowledgement

Discussions with P. Doukhan, E. Mammen and R. Rios are gratefully acknowledged.

REFERENCES

An, H-Z. and Cheng, B. (1990). A Kolmogorov-Smirnov type statistic with applications to test for normality in time series. *Int. Stat. Rev.* **59**, 287-307.

Ango Nze, P. and Doukhan, P. (1993). Estimation fonctionelle de séries temporelles mélangeantes. *C. R. Acad. Sci. Paris, Série I*, **317**, 405-408.

Ango Nze, P. and Rios, R. (1995). Estimation L^∞ de la fonction de densité d'un processus faiblement dépendant: les cas absolument régulier et fortement mélangeant. *C. R. Acad. Sci. Paris, Série I*, **320**, 1259-1262.

Ashley, R.A., Patterson, R.A. and Hinich, M.N. (1986). A diagnostic test for nonlinear serial dependence in time series fitting errors. *J. Time Ser. Anal.*, **7**, 165-78.

Bradley, R.C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*, eds. E. Eberlein & M.S. Taqqu. Birkhäuser, Boston, 165-192.

Cox, D.R. (1981). Statistical analysis of time series: some recent developments (with discussion). *Scan. J. Statist.*, **8**, 93-115.

Denker, M. and Keller, G. (1983). On U-statistics and v. Mises' statistics for weakly dependent processes. *Z. Wahr. v. Gebiete*, **64**, 505-522.

De Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probab. Theory and Related Fields*, **75**, 261-277.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statis. Assoc.*, **87**, 998-1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, **21**, 196-216.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1993). Local polynomial fitting: a standard for nonparametric regression. *Inst. Statist. Mimeo Series # 2302.*

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditonal densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83** (in press).

Franke, J., Kreis, J-P. and Mammen, E. (1995). Personal communication.

Győrfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). Non-parametric Curve Estimation from Time Series. Springer-Verlag, Berlin.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Mult. Analysis*, **14**, 1-16.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**, 1926-1947.

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

Hjellvik, V. and Tjøstheim, D. (1995). Nonparametric tests of linearity for time series. *Biometrika*, **82**, 351-68.

Hjellvik, V. and Tjøstheim, D. (1996). Nonparametric statistics for testing of linearity and serial independence. *J. Nonparametric Stat.*, **6**, 223-51.

Hjellvik, V., Yao, Q. and Tjøstheim, D. (1996). Linearity testing using local polynomial approximation. Discussion paper 60, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin, Spandauerst. 1, 10178, Berlin.

Luukkonen, R., Saikkonen, P. and Teräsvirta, T. (1988a). Testing linearity in univariate time series. *Scand. J. Statist.*, **15**, 161-75.

Poggi, J-M. and Portier, B. (1996). A test of linearity for functional autoregressive models. *J. Time Ser. Anal.*, to appear.

Rios, R. (1996). On the bias of local smoother estimators of the regression function. Preprint, University of Paris, Orsay.

Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.*, **22**, 1346 – 1370.

Shiryayev, A.N. (1984). *Probability Theorey.* New York: Springer-Verlag.

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

Skaug, H.J. and Tjøstheim D. (1993). Nonparametric tests of serial independence. In T. Subba Rao, editor, *Development in Time Series Analysis*, 207-229, Chapman and Hall.

Skaug, H.J. and Tjøstheim D. (1996). Testing for serial independence using measures of distance between densities. Springer Lecture Notes 115, P. M. Robinson and M. Rosenblatt editors, 363-77.

Yao, Q. and Tong, H. (1994). Quantifying the influence of initial values on nonlinear prediction. *J. Roy. Statis. Soc.*, **B**, **56**, 701-725.

Yoshihara, K. (1976). Limiting behaviour of U-statistics for stationary absolutely regular process. *Z. Wahr. v. Gebiete*, **35**, 237-252.

Figure captions

Figure 1a-c: The figure is based on 500 realizations of the autoregressive model (6.1) with $a = 0.5$ and $n = 100$. It shows the empirical size of $\hat{L}_T(M_1)$, $\hat{L}_T(M_1')$ and $\hat{L}_T(M_1'')$ as a function of h for $T = 0, \ldots, 4$. The numbers 0, 1, 2, 3, 4 in the plots denote the values of T. The nominal size is 0.05.

Figure 2a-c: The figure is based on 500 realizations of the exponential autoregressiv model (6.2) with $b = 1.3$ and $n = 100$. It shows the empirical power of $\hat{L}_T(M_1)$, $\hat{L}_T(M_1')$ and $\hat{L}_T(M_1'')$ as a function of h for $T = 0, \ldots, 4$. The numbers in the plots denote the values of T. The nominal size is 0.05.

Figure 3a-b: The figure is based on 500 realizations of model la) with $n = 100$. It shows the empirical power of $\hat{L}_T(V_1)$, $\hat{L}_T(V_1')$ and $\hat{L}_T(V_1'')$ as a function of h for $T = 0, \ldots, 4$. The numbers in the plots denote T. The nominal size is 0.05.

Figure 4: The figure is based on 500 realizations of the models in Table 2. It shows the power of $\hat{L}_T(M_1)$ with h cross-validated and $n = 100$, 250 and 204 for models la) - li), aa) - ag) and Aa) - Ag), respectively. The numbers 0, 1, 2, 3 denote the values of T. The + symbol indicates the power achieved in Hjellvik and Tjøstheim (1995). The nominal size is 0.05.

Figure 5: The figure is based on 500 realizations of the models in Table 2 and shows the power of $\hat{L}_T(V_1)$ with h cross-validated and $n = 100$, 250 and 204 for models la), aa) - ag) and Aa) - Ag), respectively. The numbers denote the values of T. The nominal size is 0.05.

Figure 6: The figure is based on 1000 and 20 000 (underlined symbols) realizations with $n = 100$ of model (6.1) with $a = 0$, and it shows the average empirical size of $\hat{L}_0(M_1), \ldots, \hat{L}_0(M_{10})$ as a function of the number of bootstrap replicas (m) and of the sensitivity parameter α used in the adaptive density estimate (cf. Silverman 1985, p. 100 ff) with $\alpha = 0$ corresponding to a non-adaptive density estimate. Here a, b, c, d, e and f denote nominal levels of 0.1, 0.05, 0.01, 0.001, 0.0001 and 0.00001, respectively. The bandwidth is $h = n^{-1/5}$.

Figure 7: The figure is based on 500 bootstrap replicas and the adaptive approximation to the null distribution. It shows for some real data sets the p-values for $\hat{L}_{T,\text{sup}}(M_{10})$, $\hat{L}_{T,\text{ave}}(M_{10})$, $\hat{L}_{T,\text{sup}}(V_{10})$

and $\hat{L}_{T,\text{ave}}(V_{10})$ in that order ('s' denotes 'sup' and 'a' denotes 'ave'). The numbers 0, 1, 2, 3 denote the values of T. The bandwidth is cross-validated according to Table 3, and the upper limit of the estimated order of the autoregressive fit is $\hat{p} = n/10$.

Table captions

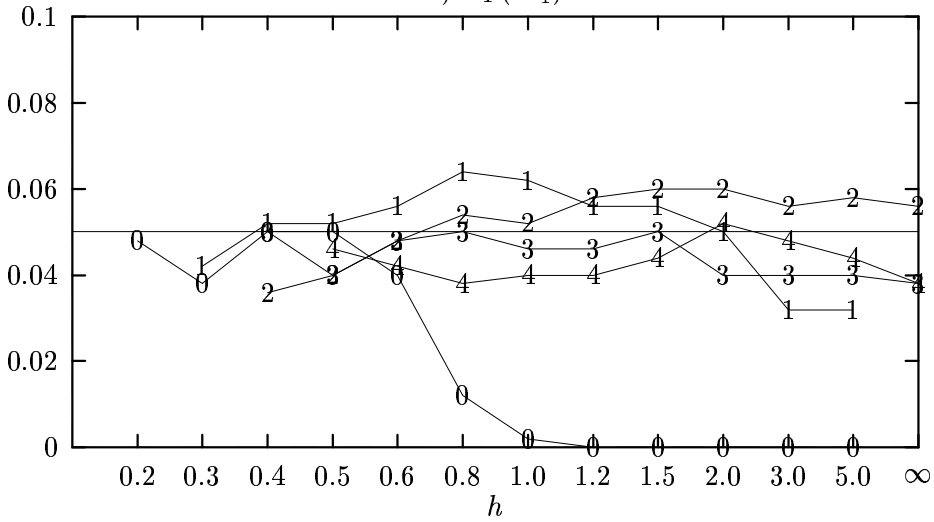
Table 1: The ratio between the asymptotic values given by Theorem 3.2 and simulated values for the mean and standard deviation of $\hat{L}_2(M_1)$, $\hat{L}_2(M_1')$ and $\hat{L}_2(M_1'')$, and the empirical sizes for these statistics when they have been centered by the asymptotic mean and scaled by the asymptotic standard deviation of Theorem 3.2. A critical value of 1.645 corresponding to a nominal size of 0.05 for the standard normal distribution has been used. The model is $X_t = \epsilon_t$, the bandwidth is $h = n^{-1/9}$ and the number of realizations are 500.

Table 2: Various nonlinear models. Models la) - lj), aa) - ag) and Aa) - Ag) are discussed in Luukkonen *et al.* (1988), Ashley *et al.* (1986) and An and Cheng (1991), respectively.

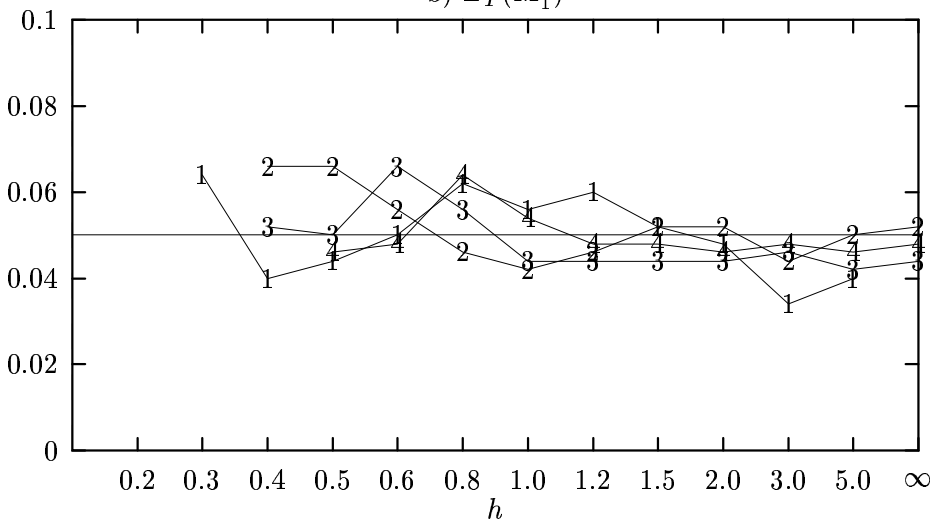
Table 3: The values among which the cross-validation routine implemented in the test chooses h. For the various tests the possible values are marked with an x.

Table 4: p-values for $\hat{L}_{0,\text{sup}}(M_{10})$, $\hat{L}_{0,\text{ave}}(M_{10})$, $\hat{L}_{0,\text{sup}}(V_{10})$ and $\hat{L}_{0,\text{ave}}(V_{10})$ for some real data sets. The table is based on 500 bootstrap replicas and the adaptive approximation to the null distribution. The data independent bandwidth $h = n^{-1/5}$ is used, and the upper limit of the estimated order of the autoregressive fit is $\hat{p} = 10$.

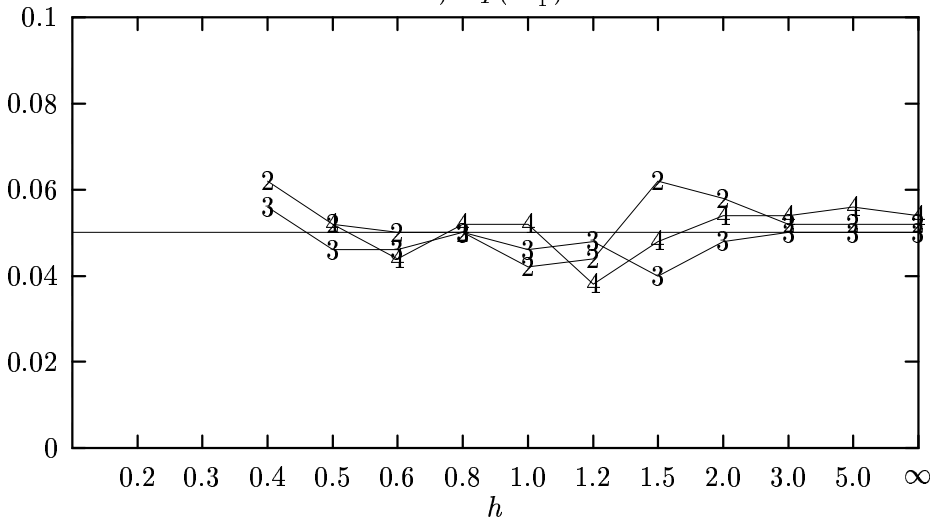
a) $\hat{L}_T(M_1)$

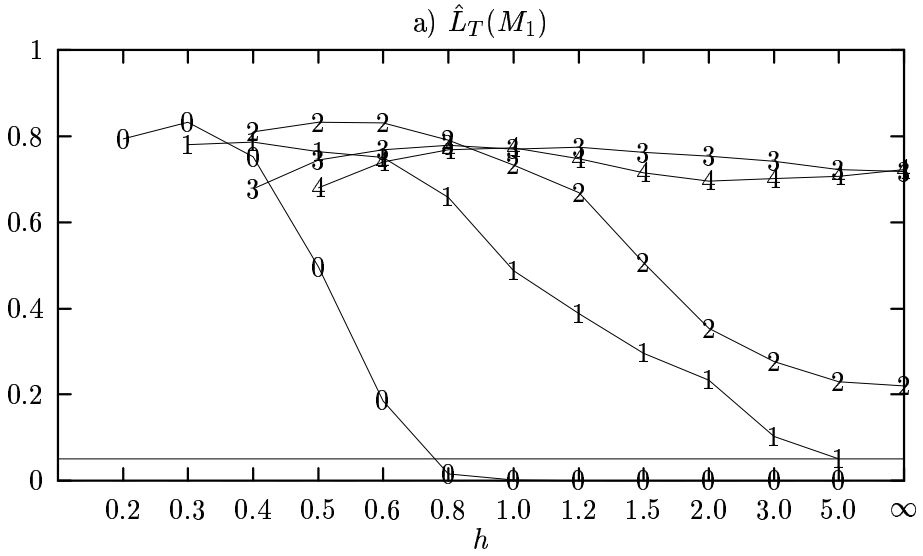


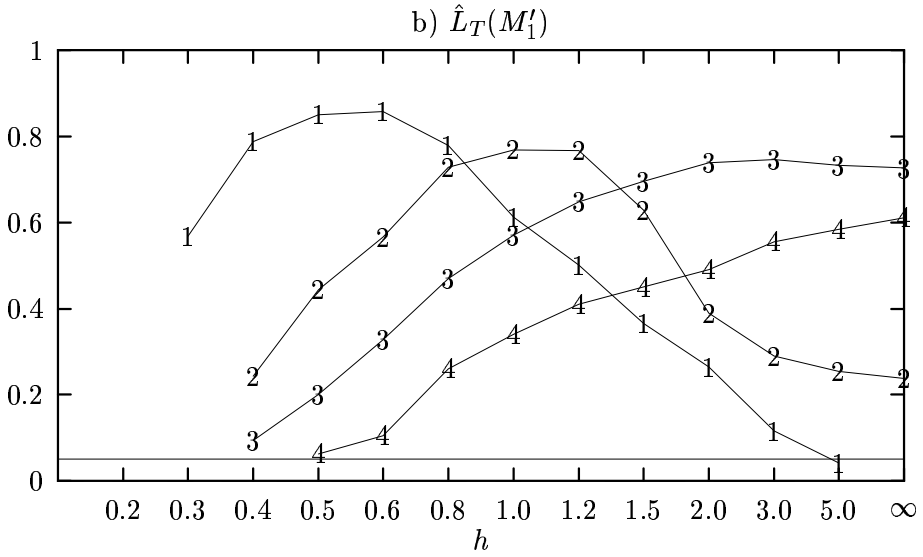
b) $\hat{L}_T(M_1')$

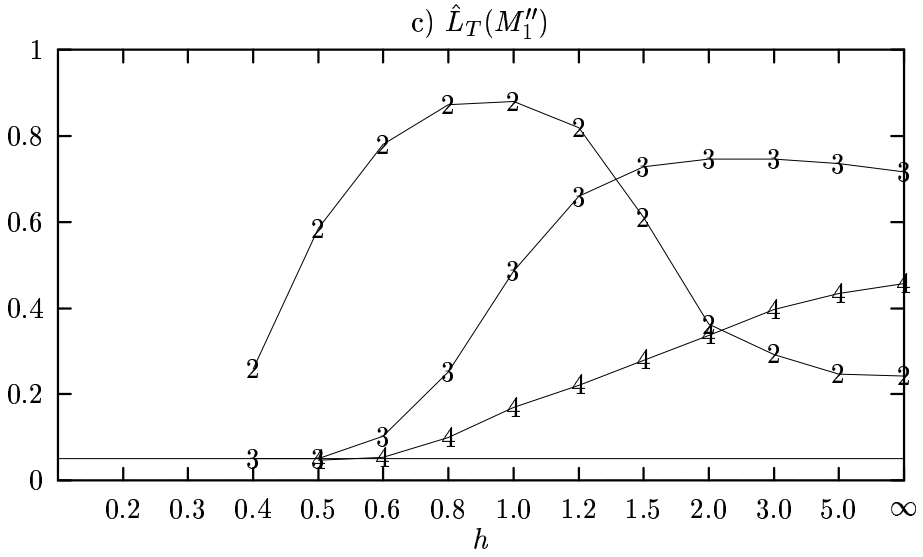


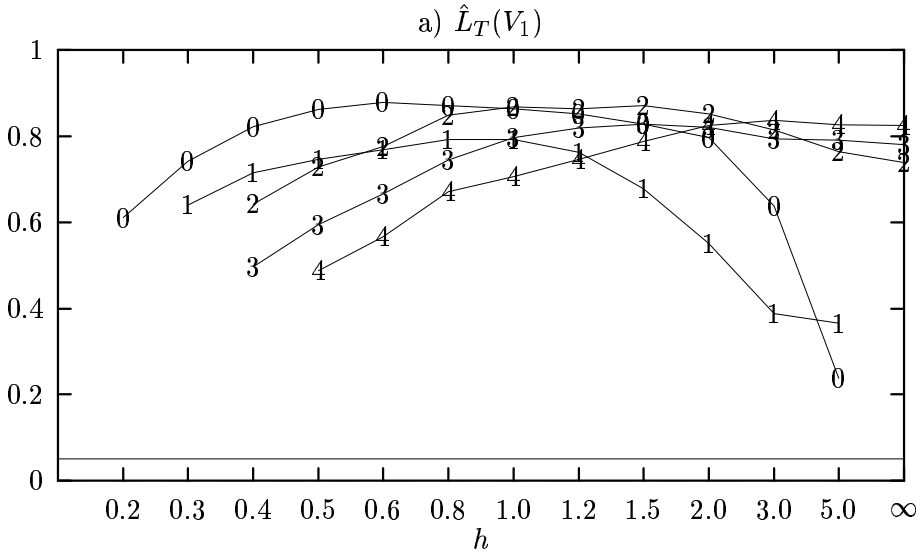
c) $\hat{L}_T(M_1'')$

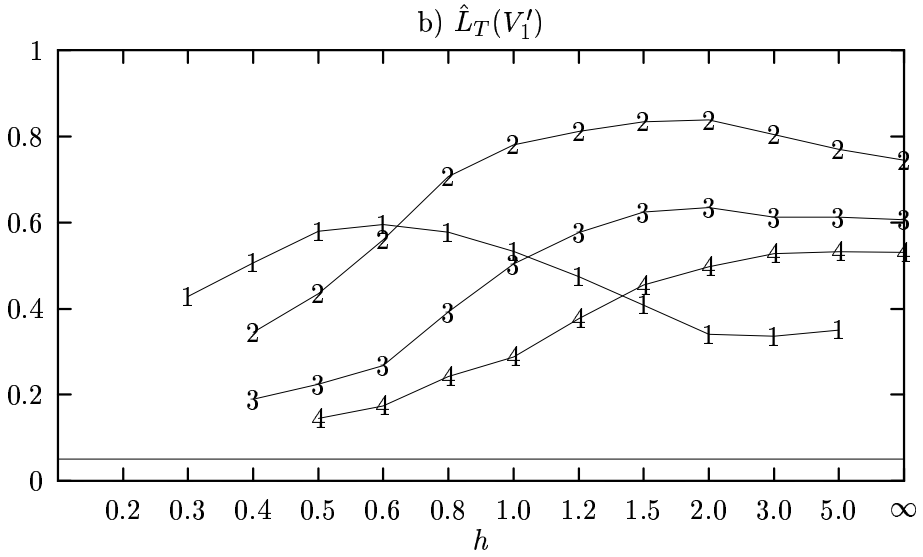


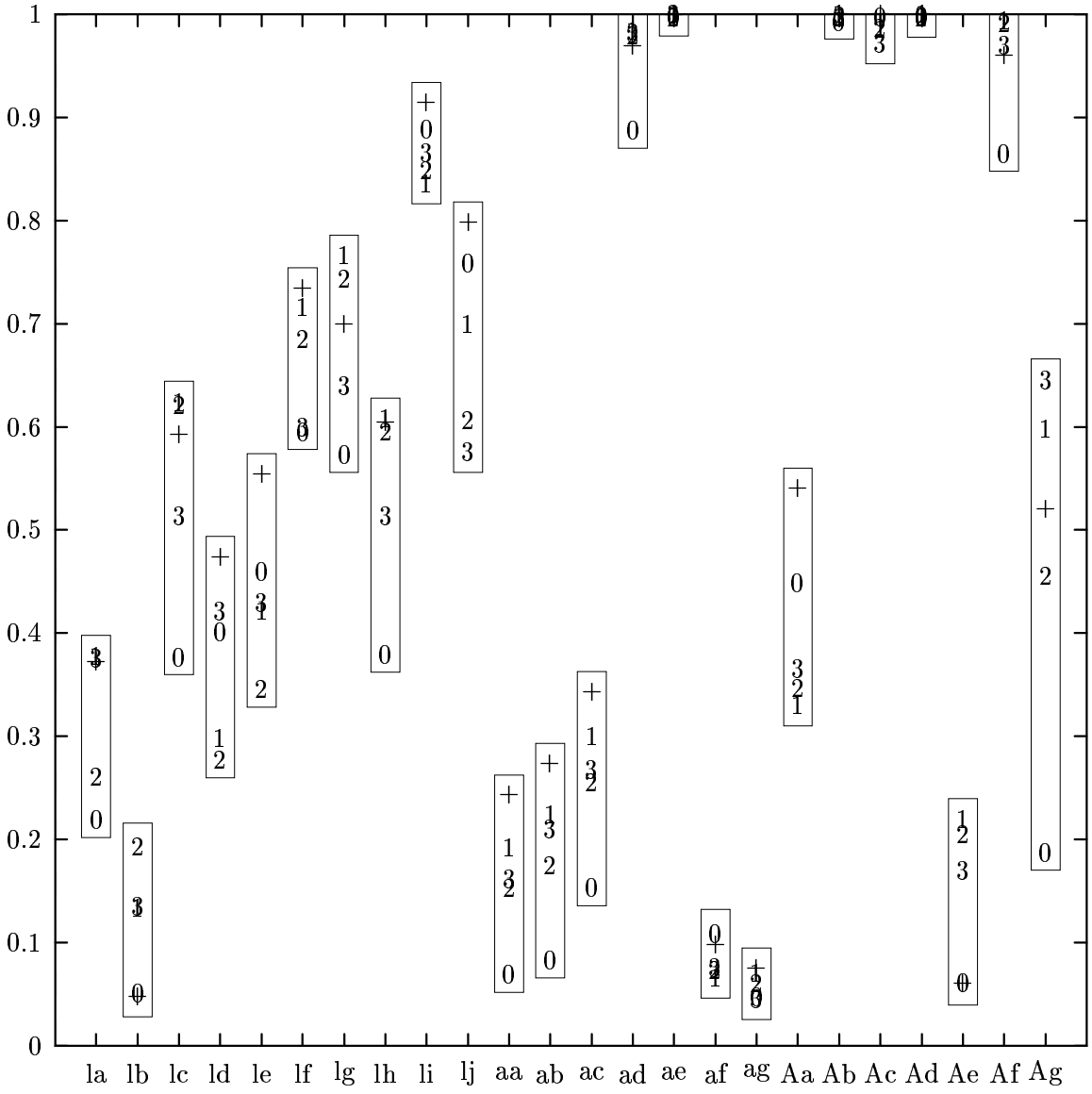

a) $\hat{L}_T(M_1)$
1
0.8
0.6
0.4
0.2
0
0.2 0.3 0.4 0.5 0.6 0.8 1.0 1.2 1.5 2.0 3.0 5.0 ∞
h

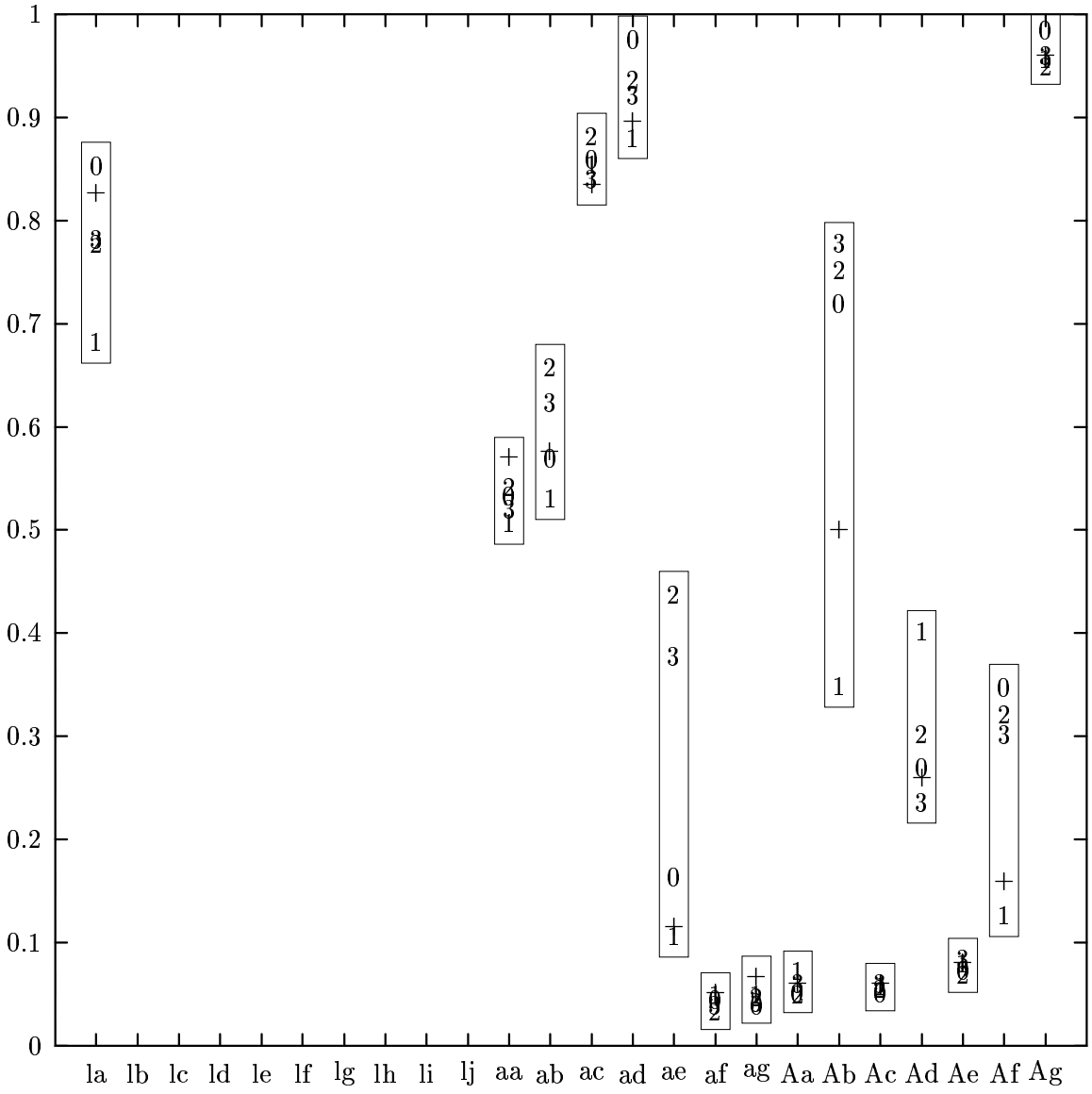

b) $\hat{L}_T(M_1')$
1
0.8
0.6
0.4
0.2
0
0.2 0.3 0.4 0.5 0.6 0.8 1.0 1.2 1.5 2.0 3.0 5.0 ∞
h


c) $\hat{L}_T(M_1'')$
1
0.8
0.6
0.4
0.2
0
0.2 0.3 0.4 0.5 0.6 0.8 1.0 1.2 1.5 2.0 3.0 5.0 ∞
h


a) $\hat{L}_T(V_1)$
1
0.8
0.6
0.4
0.2
0
0.2 0.3 0.4 0.5 0.6 0.8 1.0 1.2 1.5 2.0 3.0 5.0 ∞
h


b) $\hat{L}_T(V_1')$
1
0.8
0.6
0.4
0.2
0
0.2 0.3 0.4 0.5 0.6 0.8 1.0 1.2 1.5 2.0 3.0 5.0 ∞
h


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
la lb lc ld le lf lg lh li lj aa ab ac ad ae af ag Aa Ab Ac Ad Ae Af Ag


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
la lb lc ld le lf lg lh li lj aa ab ac ad ae af ag Aa Ab Ac Ad Ae Af Ag

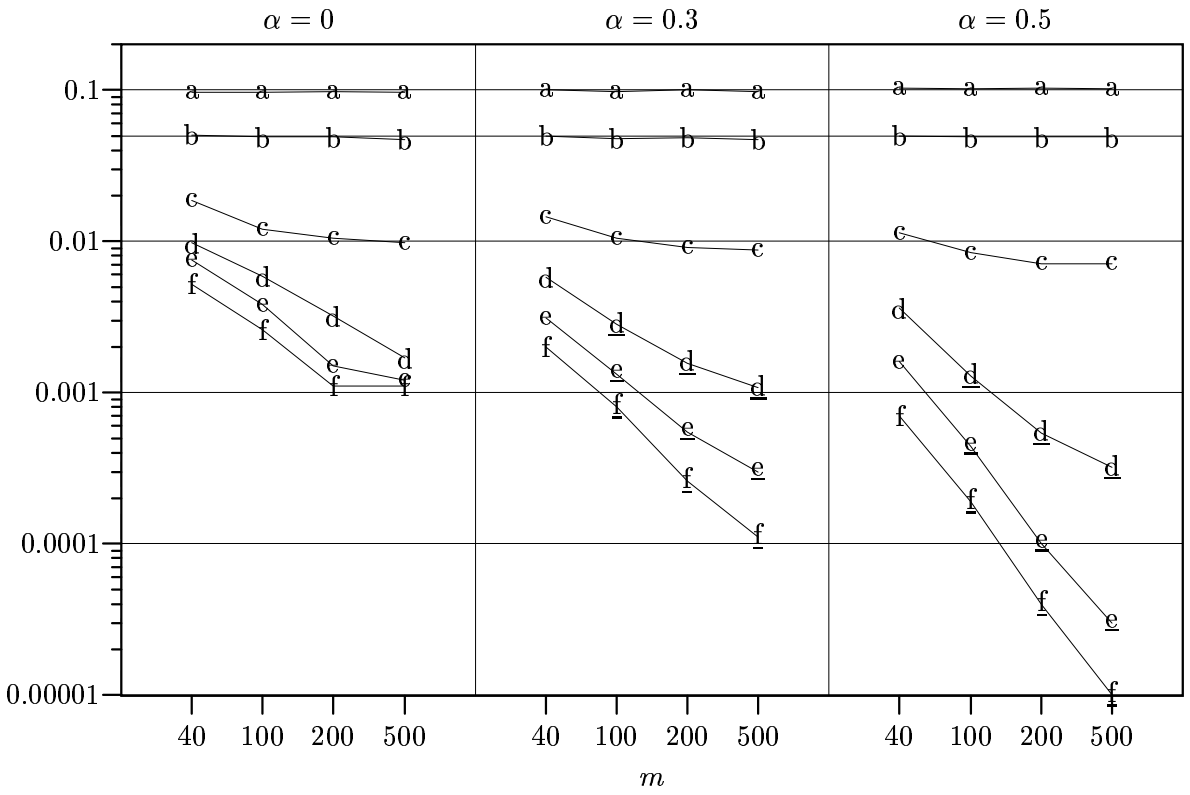


Figure 7 *p-values, real data sets.*

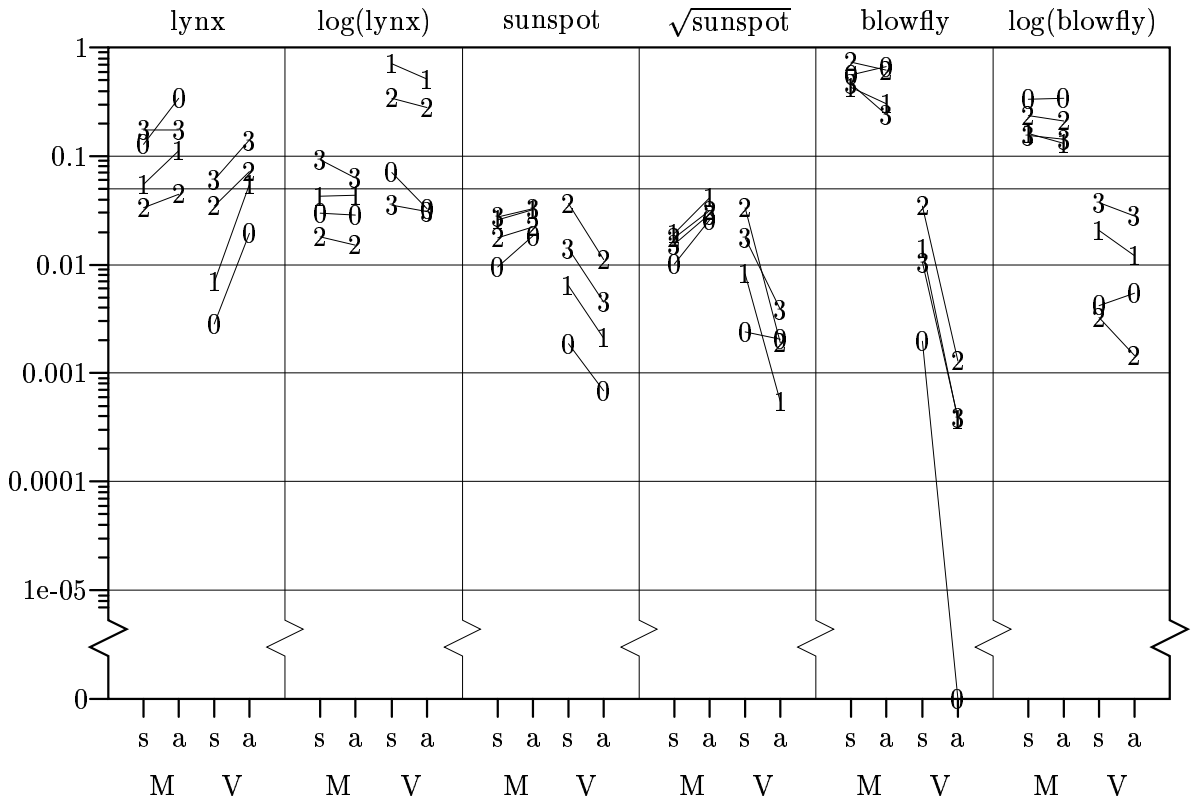


Table 1 *Asymptotic properties.*

	n:	128	256	512	2^{10}	2^{11}	2^{12}	...	2^{15}	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}	2^{21}
	mean	1.76	1.64	1.69	1.59	1.56	1.46	...	1.38	1.34	1.38	1.31	1.27	1.25	1.23
$\hat{L}_2(M_1)$	sd	1.75	1.50	1.64	1.57	1.45	1.49	...	1.40	1.28	1.30	1.27	1.17	1.17	1.20
	size	.004	.020	.012	.010	.020	.016	...	.012	.022	.020	.024	.024	.026	.030
	mean	.480	.516	.590	.620	.656	.668	...	.777	.803	.874	.886	.887	.889	.912
$\hat{L}_2(M_1')$	sd	.520	.537	.641	.684	.672	.748	...	.870	.840	.919	.937	.919	.895	.968
	size	.294	.248	.236	.214	.188	.202	...	.140	.138	.100	.092	.100	.122	.080
	mean	.423	.462	.527	.572	.616	.646	...	.760	.794	.863	.902	.901	.880	.929
$\hat{L}_2(M_1'')$	sd	.492	.481	.575	.618	.632	.732	...	.775	.811	.945	.951	.962	.883	.998
	size	.356	.300	.258	.256	.214	.222	...	.158	.140	.106	.092	.098	.122	.084

Table 2 *Nonlinear models.*

(la)	ARCH:	$X_t = 0.6X_{t-1} + \eta_t, \quad \eta_t = \epsilon_t(0.2 + 0.8\eta_{t-1}^2)^{1/2}$
(lb)	Bilinear:	$X_t = (-0.9 - 0.1\epsilon_{t-1})X_{t-1} + \epsilon_t + 2.0$
(lc)	Bilinear:	$X_t = (0.3 - 0.2\epsilon_{t-1})X_{t-1} + \epsilon_t + 1.0$
(ld)	EXPAR:	$X_t = \{0.9\exp(-X_{t-1}^2) - 0.6\}X_{t-1} + \epsilon_t$
(le)	EXPAR:	$X_t = \{0.9\exp(-X_{t-1}^2) - 0.6\}X_{t-1} + \epsilon_t + 0.3$
(lf)	EXPAR:	$X_t = \{0.9\exp(-X_{t-1}^2) - 0.6\}X_{t-1} + \epsilon_t + 1.0$
(lg)	SETAR:	$X_t = -0.3X_{t-1}1(X_{t-1} \geq 0.2) - 0.6X_{t-1}1(X_{t-1} < 0.2) + \epsilon_t$
(lh)	SETAR:	$X_t = 0.3X_{t-1}1(X_{t-1} \geq 0.2) - 0.6X_{t-1}1(X_{t-1} < 0.2) + \epsilon_t$
(li)	SETAR:	$X_t = (0.3X_{t-1} - 1.0)1(X_{t-1} \geq 0.2) - (0.3X_{t-1} + 0.5)1(X_{t-1} < 0.2) + \epsilon_t$
(lj)	SETAR:	$X_t = (0.3X_{t-1} + 1.0)1(X_{t-1} \geq 0.2) - (0.3X_{t-1} - 1.0)1(X_{t-1} < 0.2) + \epsilon_t.$
(aa)	Bilinear:	$X_t = 0.7X_{t-2}\epsilon_{t-1} + \epsilon_t$
(ab)	Nonlinear MA :	$X_t = 0.8\epsilon_{t-2}\epsilon_{t-1} + \epsilon_t$
(ac)	Extended NLMA:	$X_t = 0.8\epsilon_{t-1} + \epsilon_{t-2}\sum_{j=2}^{20}(0.8)^{j-2}\epsilon_{t-j}$
(ad)	Bilinear:	$X_t = (0.5 + 0.5\epsilon_{t-1})X_{t-1} + \epsilon_t$
(ae)	SETAR	$X_t = -0.5X_{t-1}1(X_{t-1} \leq 1) + 0.4X_{t-1}1(X_{t-1} > 1) + \epsilon_t$
(af)	Generalized SETAR:	$X_t = -(0.1 + 0.4\lvert X_{t-1}\rvert)X_{t-1}1(\lvert X_{t-1}\rvert \leq 1) - 0.5X_{t-1}1(\lvert X_{t-1}\rvert > 1) + \epsilon_t$
(ag)	EXPAR:	$X_t = \{0.9 + 0.1\exp(-X_{t-1}^2)\}X_{t-1} - \{0.2 + 0.1\exp(-X_{t-1}^2)\}X_{t-2} + \epsilon_t.$
(Aa)	EXPAR:	$\{0.3 - 0.8\exp(-X_{t-1}^2)\}X_{t-1} + \epsilon_t$
(Ab)	Bilinear:	$0.5 - 0.4X_{t-1} + 0.4X_{t-1}\epsilon_{t-1} + \epsilon_t$
(Ac)	SETAR:	$(-0.5X_{t-1} + 1)1(X_{t-1} \leq 0) + (-0.5X_{t-1} - 1)1(X_{t-1} > 0) + \epsilon_t$
(Ad)	SETAR:	$(0.5X_{t-1} + 2)1(X_{t-1} \leq 1) + (-0.4X_{t-1} + 0.5)1(X_{t-1} > 1) + \epsilon_t$
(Ae)	Nonlinear MA :	$\epsilon_t - 0.4\epsilon_{t-1} + 0.3\epsilon_{t-2} + 0.5\epsilon_t\epsilon_{t-2}$
(Af)	Nonlinear MA :	$\epsilon_t - 0.3\epsilon_{t-1} + 0.2\epsilon_{t-2} + 0.4\epsilon_{t-1}\epsilon_{t-2} - 0.25\epsilon_{t-1}^2$
(Ag)	Bilinear :	$0.4X_{t-1} - 0.3X_{t-2} + 0.5X_{t-1}\epsilon_{t-1} + 0.8\epsilon_{t-1} + \epsilon_t$

	h :	0.2	0.3	0.4	0.5	0.6	0.8	1.0	2.0	∞
$\hat{L}_T(M_1)$	$T=0$	x	x	x	x	x	x	x		
	$T=1$		x	x	x	x	x	x	x	
	$T \geq 2$			x	x	x	x	x	x	x
$\hat{L}_T(V_1)$	$T=0$		x	x	x	x	x	x	x	
	$T \geq 1$			x	x	x	x	x	x	x

Table 4 *Real data sets. Kernel estimates.*

	Blowfly	log(blowfly)	Lynx	log(lynx)	Sunspot	sqrt(sunspot)
$\hat{L}_{0,\text{sup}}(M_{10})$	.72789	.23772	.57301	.03071	.00997	.01021
$\hat{L}_{0,\text{ave}}(M_{10})$	.89241	.33802	.46900	.02568	.02028	.02586
$\hat{L}_{0,\text{sup}}(V_{10})$	.02706	.01105	.00283	.55474	.00709	.00011
$\hat{L}_{0,\text{ave}}(V_{10})$	.00107	.00053	.01907	.50939	.00451	.00272