Biometrika (1995), **82**, 2, *pp.* 369–83
Printed in Great Britain

Additivity tests for nonlinear autoregression

BY RONG CHEN
Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.

JUN S. LIU
Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.

AND RUEY S. TSAY
Graduate School of Business, University of Chicago, Chicago, Illinois 60637, U.S.A.

SUMMARY

Additivity is commonly used in the statistical literature to simplify data analysis, especially in analysis of variance and in multivariate smoothing. In this paper, we propose three procedures for testing additivity in nonlinear time series analysis. The first procedure combines some smoothing techniques with analysis of variance, the second is a Lagrange multiplier test using nonparametric estimation, and the third is a permutation test which uses smoothing techniques to obtain the test statistic and its reference distribution. We investigate properties of the proposed tests and use simulation to check their performance in finite samples. Applications of the tests to nonlinear time series analysis are discussed and illustrated by real examples.

Some key words: Alternating conditional expectation; Analysis of variance; Lagrange multiplier test; Permutation test; Tukey's one degree of freedom test.

1. INTRODUCTION

Much research in nonlinear time series analysis has focused on parametric models such as the threshold autoregressive model of Tong (1983), the exponential autoregressive model of Haggan & Ozaki (1981), and the bilinear model of Granger & Andersen (1978) and Subba Rao & Gabr (1984). Examples in the literature illustrate the usefulness of these models (Priestley, 1988; Tong, 1990). However, it is often hard to choose an appropriate class of models to entertain in a real application. To overcome this selection problem and to make use of recent advances in computing and nonparametric regression analysis, researchers have begun to use data-driven methods, such as nonparametric autoregression, in nonlinear time series analysis. Indeed, 'letting the data speak for themselves' is now a common principle for many practitioners and researchers in choosing nonlinear time series models. Nonlinear autoregression provides a general framework for such developments.

A nonlinear autoregressive model for a time series y_t is

$$y_t = f(y_{t-k_1}, \ldots, y_{t-k_p}) + \varepsilon_t, \tag{1}$$

where k_i and p are positive integers, $f(.)$ is a p-dimensional real-valued function, and ε_t is a white noise series with mean zero and finite variance (Jones, 1978). Nonparametric estimation of this model has been considered by Robinson (1983) when y_t is stationary and strong mixing. However, estimating model (1) nonparametrically often encounters

'curse of dimensionality'. Substantial amounts of data are often needed in order to obtain a variance-stable nonparametric estimate of a high-dimensional function $f(.)$. As an alternative to direct estimation of $f(.)$, we seek dimension reduction to simplify nonlinear time series analysis.

There are several tools available for dimension reduction, for example, the projection pursuit of Friedman & Stuetzle (1981) and the sliced inverse regression of Li (1991). In this paper, we use additivity to reduce the dimension. Additive models have been widely used in nonparametric regression analysis (Hastie & Tibshirani, 1991). Our choice of these models for nonlinear autoregression is based on several reasons. First, they are easier to comprehend because they do not involve interactions. Secondly, they can provide adequate approximations for many applications. Thirdly, under the additivity assumption, we can use univariate smoothing techniques directly in nonparametric estimation, resulting in a better understood estimate.

A nonlinear additive autoregressive model is of the form

$$y_t = f_1(y_{t-k_1}) + f_2(y_{t-k_2}) + \ldots + f_p(y_{t-k_p}) + \varepsilon_t, \tag{2}$$

where $f_i(.)$ are measurable, smooth functions, k_i are positive integers, and $\{\varepsilon_t\}$ is a sequence of independent and identically distributed random variables. This model was considered by Chen & Tsay (1993a). When exogenous variables are available, the model can be extended to a nonlinear additive autoregression with exogenous variables,

$$y_t = f_1(y_{t-k_1}) + \ldots + f_p(y_{t-k_p}) + g_1(x_{t-l_1}) + \ldots + g_q(x_{t-l_q}) + \varepsilon_t,$$

where x_t denotes an exogenous variable. The nonlinear additive model in (2) is analogous to the generalised additive model of Hastie & Tibshirani (1991) in regression analysis. Empirical modelling of such a time series is much easier than that of the general nonlinear model (1), needing only univariate smoothing. In addition, under additivity, the nonlinear contribution of each explanatory variable to the response variable can be easily seen; it can be displayed graphically and in some cases can be interpreted. Because additive models can be estimated more accurately than the unrestricted nonlinear autoregressive models, especially when the sample size is not large, they may provide more accurate forecasts. However, the additivity assumption is strong and should be carefully examined in practice. The goal of this paper is to provide three procedures that can be used to examine the additivity assumption in nonlinear autoregression, hence to justify the reduction from model (1) to model (2). In contrast to Auestad & Tjøstheim (1990), our procedures do not require numerical integration, nor multivariate smoothing.

The proposed testing procedures are described in §§ 2, 3 and 4, respectively. The first test uses the local conditional mean estimator of Truong (1993) and employs a procedure similar to the analysis of variance. The second test uses the alternating conditional expectation (ACE) algorithm of Breiman & Friedman (1985) to fit an additive model to the data. Additivity is then tested using a Lagrange multiplier type test. The third procedure also uses the ACE algorithm, but fits permuted residuals to some cross-product terms of the explanatory variables to obtain a reference distribution for the test statistic. Simulation is used in § 5 to illustrate the relative strength and weaknesses of the three tests. Section 6 applies the proposed tests to some real examples, and § 7 provides a brief discussion.

2. A CONDITIONAL MEAN TEST

For simplicity, we only consider nonlinear autoregressive models of order two in this section. The proposed test procedure applies equally well to higher-order models, even though the amount of computation and the sample size required grow rapidly.

For a stationary and ergodic nonlinear autoregressive time series y_t of order two, we propose a test of the additivity null hypothesis

$$H_0: y_t = f_1(y_{t-k_1}) + f_2(y_{t-k_2}) + \varepsilon_t \tag{3}$$

against the alternative $H_a: y_t = f(y_{t-k_1}, y_{t-k_2}) + \varepsilon_t$, for some k_1 and k_2. Without loss of generality, we assume $k_1 = 1$ and $k_2 = 2$.

First, a set of points $(d_1, \ldots, d_m)$ on the domain of y_t is chosen. The local conditional means of y_t on the grid points (d_i, d_j) are estimated by

$$\hat{f}(d_i, d_j) = \frac{1}{n_{ij}} \sum_{t \in I_{ij}} y_t, \tag{4}$$

where

$$I_{ij} = \{t : d_i - \delta_i/2 < y_{t-1} < d_i + \delta_i/2 \quad \text{and} \quad d_j - \delta_j/2 < y_{t-2} < d_j + \delta_j/2\},$$

n_{ij} is the number of observations in I_{ij}, and δ_i and δ_j are the length of a rectangular window around (d_i, d_j). Here we use different window sizes for different grid points to control the sample sizes in windows at different locations. Using the empirical density of the process, we can choose δ_i and δ_j so that the sample size in each cell is roughly the same. Experiment shows that the estimate in (4) works reasonably well.

Under some regular strong mixing condition, Robinson (1983) and Truong (1993) proved that, as the sample size n goes to infinity and the window sizes δ_i and δ_j go to zero,

$$(n\delta_i\delta_j)^{\frac{1}{2}}\{\hat{f}(d_i, d_j) - f(d_i, d_j)\} \sim N\{0, \sigma^2/p(d_i, d_j)\},$$

where σ^2 is the innovational variance and $p(d_i, d_j)$ is the joint density of (y_{t-1}, y_{t-2}) evaluated at (d_i, d_j). In addition, the $\hat{f}(d_i, d_j)$ are asymptotically mutually independent for all (i, j). The optimal window size δ_i is shown to be of the order $n^{-1/4}$ by Truong (1993). In practice, the value of the joint density function $p(y_{t-1}, y_{t-2})$ at the point (d_i, d_j) can be consistently estimated by $\hat{p}(d_i, d_j) = n_{ij}/(n\delta_i\delta_j)$. Hence

$$n^{\frac{1}{2}}\{\hat{f}(d_i, d_j) - f(d_i, d_j)\} \sim N(0, \sigma^2)$$

asymptotically. To test the null hypothesis H_0, we can apply the likelihood ratio test, which can be viewed as a two-way unbalanced analysis of variance by the following observation. The conditional mean estimator is based on the idea that, when the function $f(.,.)$ is sufficiently smooth, the observations in the same block have roughly the same mean values. This argument is true asymptotically as the window size goes to zero. In addition, under the strong mixing condition, the observations in a window behave like independent observations. Therefore, testing the additivity hypothesis $f(x, y) = f_1(x) + f_2(y)$ can be carried out using the conventional procedure of two-way analysis of variance.

More precisely, we partition a shrunken range, $\delta(y_{\max} - y_{\min})$, of the observations $y_1, \ldots, y_n$ into m equal intervals. Namely, we construct intervals (a_i, a_{i+1}) for $i = 0, \ldots, (m-1)$, where

$$a_i = y_{\min} + (1 - \delta)(y_{\max} - y_{\min})/2 + i\delta(y_{\max} - y_{\min})/m$$

and $\delta \in (0, 1)$ is a shrinking factor. Here we shrink the data range in order to avoid the complication of the 'boundary effect' often encountered in nonparametric smoothing procedures. Further, for simplicity in presentation, we use the same window size. For $t =$

3, ..., n, we classify y_t into the (i, j)th cell if $y_{t-1} \in (a_{i-1}, a_i)$ and $y_{t-2} \in (a_{j-1}, a_j)$ and denote such an observation by x_{ijk}, where k is used to distinguish different observations in the same cell. If y_{t-1} or y_{t-2} is outside the shrunken range, y_t is dropped from further consideration. Finally, an unbalanced two-way analysis of variance procedure is carried out to obtain an F statistic for testing the null hypothesis $H_0: f_{ij} = 0$ for all i and j in the model $x_{ijk} = \mu + \alpha_i + \beta_j + f_{ij} + \varepsilon_{ijk}$, where f_{ij} denotes a nonadditive function.

Proofs of the following theorems are given in the Appendix.

THEOREM 2·1. *Suppose the time series* (3) *is strongly mixing with a stationary density* $p(.)$ *and it satisfies the positivity condition of Besag* (1974); *i.e. under stationarity the joint density* $p(y_t, y_{t-1}) > 0$ *if and only if the marginals* $p(y_t)$ *and* $p(y_{t-1})$ *are positive. Suppose* $f_1(.)$ *and* $f_2(.)$ *are bounded continuous differentiable functions with bounded first derivatives. Then as the sample size n goes to infinity and the window size goes to zero, the test statistic F follows asymptotically, under the null hypothesis, a F-distribution with degrees of freedom* $(m-1)^2$ *and* $n^* - m^2$, *where* n^* *is the number of observations used in the two-way classification.*

The proposed testing procedure can be proven to have absolute power against local alternatives, when the sample size goes to infinity.

THEOREM 2·2. *Under the conditions of Theorem* 2·1, *the power of the test against the local alternative*

$$f(d_i, d_j) = f_1(d_i) + f_2(d_j) + g(d_i, d_j)m/n^{\frac{1}{2}} \tag{5}$$

goes to 1 *as* $n^{\frac{1}{2}}/m \to \infty$ *and* $m \to \infty$, *where* $g(d_i, d_j)$ *is a fixed nonadditive function.*

It is interesting to note that the preceding two theorems reveal a conflict between the local power and the accuracy of the F approximation to the test statistic in choosing m, the number of grid points. On one hand, m should be chosen small to increase the local power of the test. On the other hand, m must be chosen moderately large for the F approximation to be appropriate. Simulation results in § 5 reflect these observations.

For fixed grid points, Theorem 2·1 says that one can achieve consistency by shrinking the window size as the sample size increases. The optimal window size for the local conditional mean estimator is $O(n^{-1/4})$ when the main interest is in estimating the function $f(.)$ (Truong, 1993). Our result, however, reflects that it is not necessary to have an accurate knowledge of the nonparametric function $f(.)$ for the purpose of testing additivity.

Because we have an unbalanced random design and time series data are often spatially correlated in the space of (y_{t-k_1}, y_{t-k_2}), empty cells may exist, resulting in further difficulties in applications. In such cases, minor modification is needed in order to obtain correct degrees of freedom for the test statistic. One possible modification is to eliminate the entire row in which an empty cell exists. In addition, the reference row should be chosen carefully. Typically, cells in the centre row and column have more observations and may be chosen as the reference row and column.

3. A LAGRANGE MULTIPLIER TEST

Nonlinear additive models have been studied extensively for independent data, and many modelling procedures suggested. In particular, back-fitting algorithms such as the alternating conditional expectation algorithm of Breiman & Friedman (1985) and the additivity and variance stabilisation algorithm of Tibshirani (1988) are easy to use and

have nice properties. Using these tools, we can build additivity tests similar to Tukey's one degree of freedom test (Tukey, 1949) based on the goodness of fit of an additive model.

Using the Volterra expansion, we can rewrite the nonlinear autoregressive model $y_t = f(y_{t-k_1}, \ldots, y_{t-k_p}) + \varepsilon_t$ as

$$\begin{aligned} y_t &= \mu + \sum_{u=1}^{p} \phi_u y_{t-k_u} + \sum_{u \leqslant v}^{p} \phi_{uv} y_{t-k_u} y_{t-k_v} + \sum_{u \leqslant v \leqslant w}^{p} \phi_{uvw} y_{t-k_u} y_{t-k_v} y_{t-k_w} + \ldots + \varepsilon_t \\ &= \mu + \sum_{u=1}^{p} f_u(y_{t-k_u}) + \sum_{u<v}^{p} \phi_{uv} y_{t-k_u} y_{t-k_v} + \sum_{u \leqslant v \leqslant w}^{p} \phi_{uvw} y_{t-k_u} y_{t-k_v} y_{t-k_w} + \ldots + \varepsilon_t, \quad (6) \end{aligned}$$

where

$$f_u(y_{t-k_u}) = \phi_u y_{t-k_u} + \phi_{uu} y_{t-k_u}^2 + \phi_{uuu} y_{t-k_u}^3 + \ldots$$

is the collection of all the powers of y_{t-k_u} and it is understood that $u = v = w$ is excluded from the summation of the third-order term. If the model is additive, all the coefficients of the higher-order terms in the preceding equation should be zero. Based on this observation, we propose the following three-step test.

1. An additive model $y_t = f_1(y_{t-k_1}) + \ldots + f_p(y_{t-k_p}) + \varepsilon_t$ is estimated using the ACE algorithm with the restriction that the response variable can only be linearly transformed. Denote the estimates of $f_i(.)$ by $\hat{f}_i(.)$ and the residuals by $\hat{\varepsilon}_t = y_t - \sum_{i=1}^{p} \hat{f}_i(y_{t-k_i})$.
2. Regress the cross-product terms $y_{t-k_{j_1}} y_{t-k_{j_2}}$ on $y_{t-k_1}, \ldots, y_{t-k_p}$ for $1 \leqslant i_1 < i_2 \leqslant p$, and the third-order cross-product terms $y_{t-k_{j_1}} y_{t-k_{j_2}} y_{t-k_{j_3}}$ on $y_{t-k_1}, \ldots, y_{t-k_p}$ for $1 \leqslant j_1 \leqslant j_2 \leqslant j_3 \leqslant p$ except for $j_1 = j_2 = j_3$ using the ACE algorithm. This procedure results in

$$K = p(p-1)/2 + \{p(p+1)(p+2)/6 - p\} = (p-1)p(p+7)/6$$

   residual series, say, $e_1(t), \ldots, e_K(t)$. Here the transformations of the response variables are also restricted to be linear.
3. Linearly regress the residual series $\hat{\varepsilon}_t$ obtained from step 1 on $e_1(t), \ldots, e_K(t)$ obtained from step 2. Compute the test statistic nR^2 where n is the sample size and R^2 is the conventional coefficient of determination in linear regression analysis.

In applications, the orders p and k_i can be chosen in several ways. For example, one can use the nonparametric procedure of Tjøstheim & Auestad (1994) to select significant lagged variables. Another method is to let $k_i = i$ and to use a stepwise procedure in choosing p, starting with $p = 1$. In step 2, the order of cross-product terms is not limited to three. Higher-order cross-product terms can be included to increase the power of the test if the sample size is sufficiently large.

To obtain the asymptotic distribution of the proposed Lagrange multiplier test, we define the optimal transformations as functions $f_1^*, \ldots, f_p^*$ that minimise

$$e^2(\tilde{f}_1, \ldots, \tilde{f}_p) = E\left\{ y_t - c - \sum_{j=1}^{p} \tilde{f}_i(y_{t-k_j}) \right\}^2$$

among all $\tilde{f}_i(.)$ such that $E\{\tilde{f}_i(y_t)\} = 0$. This definition is slightly different from that of Breiman & Friedman (1985) because we do not require any transformation of the response variable y_t. Breiman & Friedman (1985) showed that, under some mild condition, the optimal transformation exists. However, those optimal transformations may differ from

the functions $f_i(.)$ of the regression model $y_t = f(y_{t-k_1}) + \ldots + f(y_{t-k_p}) + \varepsilon_t$. Nevertheless, it is easy to prove, following the argument of Theorem 5.2 of Breiman & Friedman (1985), that, if the response variable is not transformed, then $c + \sum_{i=1}^{p} f_i^*(.) = \sum_{i=1}^{p} f_i(.)$ in our case.

LEMMA 3·1. *Assume y_t is a stationary ergodic time series satisfying*

$$y_t = f_1(y_{t-k_1}) + \ldots + f_p(y_{t-k_p}) + \varepsilon_t,$$

where ε_t are independent and identical distributed white noise with zero mean and finite fourth moment. The marginal density of y_t is assumed to be positive on $\Re^1$. The functions $f_1(.), \ldots, f_p(.)$ and the joint density of $(y_t, y_{t-k_1}, \ldots, y_{t-k_p})$ satisfy the necessary conditions for the ACE algorithm to converge. In addition, we assume that the optimal transformations f_i^* are continuous. The smoothers used in the ACE algorithm are assumed to be linear, uniformly bounded and uniformly consistent on any fixed compact set as the sample size n goes to infinity. Then, for any fixed compact set, the estimate $f_i^{(n)}$ obtained by the ACE algorithm converges uniformly to the optimal transformation f_i^* for $i = 1, \ldots, p$ with probability one.

The lemma can be proven in the same way as Theorem A.2 of Breiman & Friedman (1985) or Theorem 2.1 of Koyak (1990) with some minor modification, and hence is omitted.

THEOREM 3·1. *Under the conditions of Lemma 3·1, if the three-step testing procedure is applied on a compact set of $\Re^1$, by discarding observations outside the compact set, the test statistic nR^2 is asymptotically chi-squared with $(p-1)p(p+7)/6$ degrees of freedom.*

Proof. Since $\hat{f}_i^{(n)} \to f_i^*$ uniformly and almost surely on the compact set, and $c + \sum f_i^*(y_{t-k_i}) = \sum f_i(y_{t-k_i})$, Slutsky's theorem ensures that $\hat{\varepsilon}_t$ follows the same distribution as ε_t asymptotically. The series $e_i(t)$ converges almost surely to the residual series $e_i^*(t)$ computed using the true optimal transformations at the observed data. Note that $e_i^*(t)$ depends only on y_{t-i} for $i > 0$. Hence ε_t and $e^*(t)$ are uncorrelated. The rest of the proof follows the same lines as that of Theorem 1 of Tsay (1986) or Lemma 3.1 of Keenan (1985). □

Remark 1. The optimal transformations obtained by the ACE algorithm are generally not the transformations in the regression model (Breiman & Friedman, 1985). However, if the response variable is not transformed, the optimal transformations estimated by the ACE algorithm are the regression functions.

Remark 2. The continuity of the optimal transformations depends solely on the joint stationary distribution of the time series, and is usually difficult to check in practice. We conjuncture that if the functions f_i are continuous, the optimal transformations will be continuous. This is, again, due to the fact that the response variable is not being transformed.

Remark 3. The requirement that the smoother used be uniformly consistent on a compact set is a mild condition. For example, a direct application of Lemma A.1 of Breiman & Friedman (1985) shows that the nearest neighbour smoother is uniformly consistent on a compact set.

4. A PERMUTATION TEST OF ADDITIVITY

If we also group the high power terms in the Volterra expansion in (6), e.g. letting

$$f_{uv}(y_{t-k_u}y_{t-k_v}) = \phi_{uv}y_{t-k_u}y_{t-k_v} + \phi_{uuvv}(y_{t-k_u}y_{t-k_v})^2 + \phi_{uuuvvv}(y_{t-k_u}y_{t-k_v})^3 + \ldots,$$

then we have

$$y_t = \mu + \sum_{u=1}^{p} f_u(y_{t-k_u}) + \sum_{u<v} f_{uv}(y_{t-k_u}y_{t-k_v}) + \sum_{u\leqslant v\leqslant w} f_{uvw}(y_{t-k_u}y_{t-k_v}y_{t-k_w}) + \ldots + \varepsilon_t.$$

It is, therefore, of interest to test the null hypothesis

$$H_0: y_t = \mu + \sum_{u=1}^{p} f_u(y_{t-k_u}) + \varepsilon_t$$

versus the alternative

$$H_\alpha: y_t = \mu + \sum_{u=1}^{p} f_u(y_{t-k_u}) + \sum_{u<v}^{p} f_{uv}(y_{t-k_u}y_{t-k_v}) + \varepsilon_t.$$

Although such a test is limited to the first order cross-product terms, it should have decent power against a large class of nonadditive models. Here the usual F-ratio test is not directly applicable because we use nonparametric tools such as the ACE algorithm in model fitting, and hence it is difficult to compute the appropriate degrees of freedom. Instead we propose a test based on simple permutations:

1. As for step 1 of § 3.
2. Regress the estimated residuals $\hat{\varepsilon}_t$ from step 1 on the cross-product terms $y_{t-k_i}y_{t-k_j}$ for $1 \leqslant i < j \leqslant p$ using the ACE algorithm and obtain the sum of squares of residuals of this regression.
3. Form a new series of residuals $e(t)$ by permuting the $\hat{\varepsilon}_t$. Regress it on the same cross-product terms as those of step 2 using the ACE algorithm and obtain the sum of squares of residuals. Repeat this step N times.
4. The p-value of the permutation test is determined by the proportion of the sum of squares of residuals in step 3 that is smaller than sum of squares of residuals in step 2.

The proposed permutation test is based on the fact that, under the null hypothesis, the residuals $\hat{\varepsilon}_t$ are asymptotically independent and identically distributed. Hence any permutation of the $\hat{\varepsilon}_t$ should have the same joint distribution as that of $\hat{\varepsilon}_t$. Further, when these residuals are regressed on the cross-product terms using the ACE algorithm, the residual sum of squares for the original $\hat{\varepsilon}_t$ should also have the same distribution as that for the permuted ones. On the other hand, if the model is indeed nonadditive, a random permutation of the $\hat{\varepsilon}_t$ loses the information of the cross-product terms $f_{uv}(y_{t-k_u}y_{t-k_v})$ contained in $\hat{\varepsilon}_t$, resulting in a larger sum of squares of residuals.

5. SIMULATION STUDY

5·1. *Models used in the study*

In this section, we demonstrate the performance of the proposed three tests by simulation. Two sets of models are used; one for size consideration and the other for power

study. The first set consists of four additive models:

$$y_t = 0{\cdot}8y_{t-1} - 0{\cdot}3y_{t-2} + \varepsilon_t, \tag{7}$$

$$y_t = 0{\cdot}5y_{t-1} + \sin(y_{t-2}) + \varepsilon_t, \tag{8}$$

$$y_t = 2\exp(-0{\cdot}1y_{t-1}^2)y_{t-1} - \exp(-0{\cdot}1y_{t-2}^2)y_{t-2} + \varepsilon_t, \tag{9}$$

$$y_t = -2y_{t-1}I(y_{t-1} \leqslant 0) + 0{\cdot}4y_{t-1}I(y_{t-1} > 0) + \varepsilon_t, \tag{10}$$

where $I(x)$ is an indicator variable such that $I(x) = 1$ if x holds. These models are used to study the behaviour of the proposed tests under the null hypothesis. They represent time series models commonly used in univariate analysis. For example, model (10) is a threshold autoregressive model with an asymmetric limit cycle, and model (8) contains a sine function at lag 2. Trigonometric functions have been used in the time series literature to describe periodic series (Lewis & Ray, 1993). Model (9) is similar to an exponential autoregressive model, but uses different lag variables in the exponents so that the model is additive. We chose the linear AR (2) model of (7) to ensure that the proposed tests work well for this simple case. The threshold model is also used to check the performance of the tests when the order is misspecified.

The second simulation considers four nonadditive models:

$$y_t = 2\exp(-0{\cdot}1y_{t-1}^2)y_{t-1} - \exp(-0{\cdot}1y_{t-1}^2)y_{t-2} + \varepsilon_t, \tag{11}$$

$$y_t = y_{t-1}\sin(y_{t-2}) + \varepsilon_t, \tag{12}$$

$$y_t = (0{\cdot}5y_{t-1} - 0{\cdot}4y_{t-2})I(y_{t-1} < 0) + (0{\cdot}5y_{t-1} + 0{\cdot}3y_{t-2})I(y_{t-1} \geqslant 0) + \varepsilon_t, \tag{13}$$

$$y_t = 14{\cdot}27 + 0{\cdot}46y_{t-1} - 0{\cdot}02y_{t-1}(y_{t-2} - 30)_+ + 0{\cdot}047y_{t-1}(30 - y_{t-2})_+ + \varepsilon_t, \tag{14}$$

where $(x)_+ = x$ if $x > 0$, and $(x)_+ = 0$ otherwise. These models are used to study the power of the proposed tests. They are either models commonly used in the literature or models that were suggested for real time series. For example, model (12) is a functional-coefficient AR (1) model with a sine function of lag 2 (Chen & Tsay, 1993b), and model (14), which exhibits limiting cycle behaviour, is an adaptive spline threshold autoregression: see Lewis & Stevens (1991), who show that such models are useful in real applications such as modelling and forecasting the annual sunspot numbers.

5·2. *Simulation results*

For each of the models in (7)–(14), we applied the proposed three tests to 300 realisations, each with 300 observations. The sample size of 300 or larger is common in nonlinear time series analysis, especially in using nonparametric methods. It is often difficult to obtain a reliable estimate of the high-dimensional surface when the sample size is small. The innovations ε_t used are independent standard normal random variates. In applying the conditional mean test of § 2, we used a shrinking factor $\delta = 0{\cdot}8$ and a 5×5 grid. For each realisation, we computed the F-statistic and its p-value with respect to the corresponding asymptotic F-distribution. Table 1(a) shows the percentiles of the asymptotic p-values for models (7)–(10). From the table, we see that the asymptotic p-values are close to the uniform distribution in the unit interval $[0, 1]$, indicating that the proposed conditional mean test follows closely the asymptotic F-distribution of Theorem 2·1 when the null hypothesis holds. Table 2(a) shows the percentages of rejection by the conditional mean test under different significance levels for models (11)–(14). The test has good power

Table 1. *Percentiles of p-values of the proposed additivity tests under the null hypothesis*

(*a*) *The conditional mean test*

Probability	Model (7)	Model (8)	Model (9)	Model (10)
0·01	0·016	0·007	0·016	0·017
0·05	0·040	0·045	0·051	0·052
0·10	0·091	0·105	0·119	0·098
0·25	0·282	0·219	0·295	0·240
0·50	0·507	0·427	0·574	0·507
0·75	0·748	0·696	0·802	0·748
0·90	0·897	0·886	0·915	0·919
0·95	0·946	0·928	0·955	0·960
0·99	0·996	0·982	0·990	0·993

(*b*) *The Lagrange multiplier test*

Probability	Model (7)	Model (8)	Model (9)	Model (10)
0·01	0·010	0·013	0·002	0·026
0·05	0·041	0·046	0·032	0·091
0·10	0·099	0·092	0·085	0·158
0·25	0·271	0·232	0·230	0·315
0·50	0·549	0·497	0·445	0·492
0·75	0·791	0·763	0·739	0·764
0·90	0·910	0·914	0·907	0·912
0·95	0·959	0·951	0·951	0·955
0·99	0·996	0·986	0·976	0·996

(*c*) *The permutation test*

Probability	Model (7)	Model (8)	Model (9)	Model (10)
0·01	0·025	0·010	0·035	0·060
0·05	0·090	0·075	0·100	0·160
0·10	0·175	0·160	0·185	0·250
0·25	0·360	0·365	0·360	0·520
0·50	0·660	0·675	0·665	0·835
0·75	0·865	0·850	0·890	0·950
0·90	0·960	0·950	0·960	0·980
0·95	0·990	0·960	0·980	0·990
0·99	1·000	1·000	1·000	1·000

against threshold models and some functional-coefficient autoregressive models. However, it has low power against the exponential model in (11).

Because the sample size used is 300 for each realisation, we added the term $y_{t-1}^2 y_{t-2}^2$ to step 2 of the proposed Lagrange multiplier test to increase the power of the test. This results in using an asymptotic chi-squared distribution with 4 degrees of freedom for the test statistic. Table 1(b) gives the p-values of the test statistics with respect to χ_4^2 for models (7)–(10). Again, the proposed Lagrange multiplier test follows closely its limiting distribution under the null hypothesis. Table 2(b) gives the power of the Lagrange multiplier test under the alternative hypothesis. For the models considered, the test has good power.

Finally, Table 1(c) shows the p-values of the proposed permutation test of § 4 for models (7)–(10). For ease in computation, we used 100 permutations for each realisation and only considered the cross-product term $y_{t-1}y_{t-2}$ in step 3 of the proposed test. Increasing

Table 2. *Percentages of rejection by the proposed tests under the alternative hypothesis*

(*a*) *The conditional mean test*

Significance level (%)	Model (11)	Model (12)	Model (13)	Model (14)
10	0·403	1·000	0·740	0·873
5	0·267	1·000	0·627	0·817
1	0·117	1·000	0·450	0·597

(*b*) *The Lagrange multiplier test*

Significance level (%)	Model (11)	Model (12)	Model (13)	Model (14)
10	1·000	0·990	0·943	1·000
5	0·997	0·990	0·897	1·000
1	0·977	0·967	0·723	1·000

(*c*) *The permutation test*

Significance level (%)	Model (11)	Model (12)	Model (13)	Model (14)
10	0·090	1·000	0·780	0·903
5	0·037	1·000	0·647	0·843
1	0·007	1·000	0·260	0·587

the number of permutations and the number of cross-product terms would substantially increase the computing time. From the table, the p-values appear to be larger than for a uniform distribution in [0, 1], resulting in smaller type-I errors. This might be due to the fact that the ACE algorithm used tends to overfit the data in the finite sample case. Table 2(c) gives the percentages of rejection by the permutation test for models (11)–(14). The test has reasonable power against models (12)–(14), but fares poorly against model (11). This latter behaviour will be discussed later.

5·2. *Some discussion*

The poor performance of the permutation test against model (11) is understandable, because the nonadditivity of the model is in higher-order terms and we only used the simple cross-product term $y_{t-1}y_{t-2}$ in the test. We reran the test using the four cross-product terms $y_{t-1}y_{t-2}$, $y_{t-1}^2y_{t-2}$, $y_{t-1}y_{t-2}^2$ and $y_{t-1}^2y_{t-2}^2$. The percentages of rejection by the test increase to 0·48, 0·36 and 0·17, respectively, for the significance level 10%, 5% and 1%. In practice, it might be helpful to employ several cross-product terms in using the permutation test.

Secondly, the permutation test encounters some difficulty in testing model (10), suggesting that the test might be sensitive to model misspecification. On the other hand, the conditional mean test works well for model (10).

Thirdly, to study the 'bandwidth' effect on the conditional mean test, we simulated 300 series, each with 500 observations, for each of the models in (11)–(14) and performed the test using 5×5 and 7×7 grid points. The percentages of rejecting additivity are shown in Table 3. Clearly, the choice of window size affects the power of the test, but not in a unidirectional manner. The effects depend on the underlying functions of the model. For threshold models which have discontinuity points, increasing the grid points, i.e. reducing the window size, actually reduces the power of the test. Further study is needed to choose the optimal window size in using the conditional mean test. Finally, comparing Table 2(a)

Table 3. *Power comparison of the conditional mean test for different window sizes*

Significance level (%)	Model (11)		Model (12)		Model (13)		Model (14)	
	5×5	7×7	5×5	7×7	5×5	7×7	5×5	7×7
10	0·69	0·83	1·00	1·00	0·93	0·86	0·98	0·98
5	0·57	0·75	1·00	1·00	0·89	0·80	0·97	0·98
1	0·33	0·55	1·00	1·00	0·72	0·62	0·91	0·94

and Table 3, we see that, as expected, increasing the sample size improves the power of the tests.

6. A REAL EXAMPLE

In this section we apply the proposed Lagrange multiplier test to the daily sea surface temperature at Granite Canyon on the California coast. The data, in degrees centigrade, were from 1971 to 1991 for 7361 observations. The objective of this application is to illustrate that the proposed tests can be used to validate previous analyses available in the literature and to reduce the dimension in model building.

Following Lewis & Ray (1993), we considered the logarithms of the data. First, using $p = 4$ and $k_i = i$, the Lagrange multiplier test gives a p-value of 0·03 for the null hypothesis of additivity, implying that interactions exist between the lagged values of the log sea surface temperature. This result is in agreement with the study of Lewis & Ray (1993).

Secondly, we considered a log salinity series as an exogenous variable and employed data from April 1986 to March 1991 only for 1915 observations. Figure 1 shows the log temperature and log salinity series. Lewis & Ray (1993) further included wind speed and wind direction as exogenous variables. However, because of their categorical nature, wind speed and direction were dropped from our analysis. Using the first four lagged variables of y_t and the first four exogenous variables x_{t-i} for $i = 0, \ldots, 3$, we obtained a p-value of 0·134 for the Lagrange multiplier test. Thus, the additivity assumption appears to be reasonable for the daily sea surface temperature data in the presence of salinity. Further study found that an additive model that includes lagged temperature y_{t-i} with $i = 1, 2, 4$ and lagged salinity series x_{t-i} for $i = 0, 1, 3$ fits the data well. Figure 2 shows the fitted values and the residuals. The estimated residual standard error of the model is 0·038 which is very close to 0·039 of Lewis & Ray (1993) who used a nonadditive model that employs lagged temperature y_{t-i} for $i = 1, 2, 8, 34, 39$, lagged wind speed w_{t-i} for $i = 1, 32, 44$, and two three-way interactions between y_{t-i} and w_{t-j}. One of the three-way interactions is

$$(2{\cdot}25 - y_{t-34})_+ (y_{t-39} - 2{\cdot}29)_+ (2{\cdot}49 - w_{t-1})_+,$$

which is hard to interpret.

Because of the large sample size and the number of lagged variables involved, we did not use the conditional mean and permutation tests in this particular example. These two tests require intensive computation. For this reason, they may encounter some limitations in real applications.

7. DISCUSSION

We have shown that the conditional mean test is reliable under the null hypothesis. However, it is practical only when the order of the autoregression is low. When the

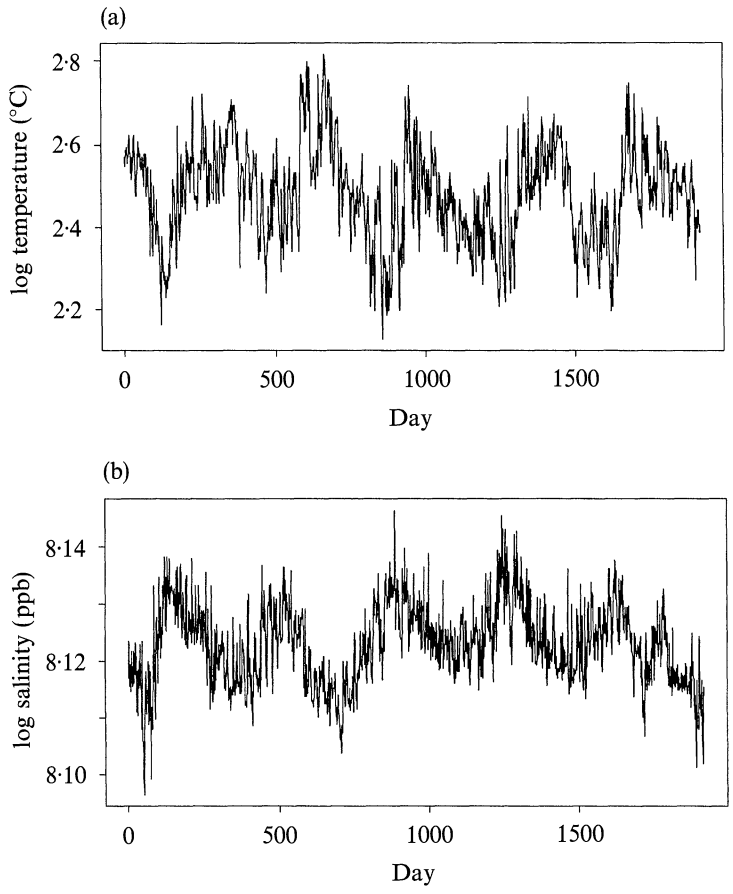

Fig. 1. Sea temperature and salinity data from April 1986 to March 1991. (a) Log daily sea temperature in degrees centrigrade. (b) Log daily salinity in parts of solute per 1000 milligrams of water (ppb).

autoregressive order is high, the computation involved and the sample size required grow exponentially. The Lagrange multiplier test is shown to be powerful. It can easily be implemented, even for higher-order models, because it only involves one-dimensional smoothing. However, the alternative models of the test are relatively limited and the test could be sensitive to model misspecification. The permutation test is computationally demanding, but it does not require any strong assumption on the innovations. Consequently, the choice between the three tests in a real application would depend on the particular circumstance. For example, if substantive prior information is available to suggest possible nonlinear relation between the variables, then the Lagrange multiplier test seems reasonable.

ACKNOWLEDGEMENT

This research is supported in part by the National Science Foundation and by the Graduate School of Business, University of Chicago. We thank Bonnie Ray for the daily sea temperature data and Wlodzimierz Bryc for valuable comments on the optimal transformations of the ACE algorithm. We also thank the Editor, Associate Editor and the referee for useful comments and suggestions.

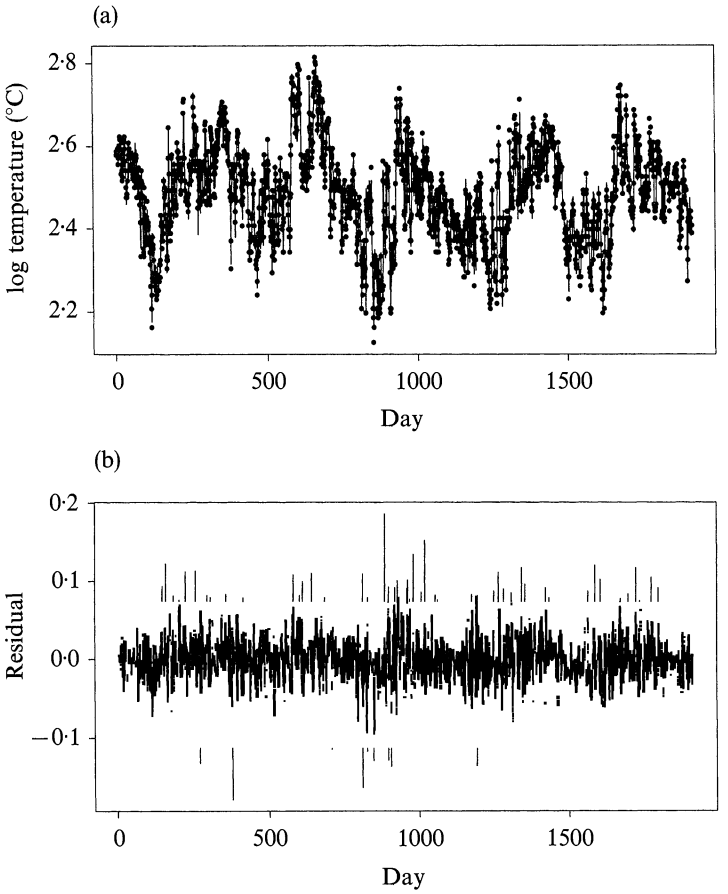

Fig. 2. Data of Fig. 1. (a) Fitted plot; dots are observed values and solid lines connect the fitted values. (b) Residual plot.

Appendix

Proofs of Theorems 2·1 and 2·2

Proof of Theorem 2·1. If the support of $p(.)$ is not a closed interval, we can find an 'action interval' $[a, b]$ such that $p(x) > 0$ on $[a, b]$ and $p([a, b]) > \eta$ for some $\eta > 0$. Denote by n^* the number of observations satisfying $(y_{t-1}, y_{t-2}) \in [a, b] \times [a, b]$. Clearly, $n^* = O(n)$ and $n^* \to \infty$ as $n \to \infty$. Let $a = a_0 < a_1 < \ldots < a_m = b$ partition the interval $[a, b]$ into m subintervals. Hence (a_i, a_j) for $i, j = 0, \ldots, m$ are $(m+1)^2$ grid points on the plane. Denote by c_{ij} the small square with vertices (a_{i-1}, a_{j-1}), (a_i, a_{j-1}), (a_{i-1}, a_j) and (a_i, a_j), and by n_{ij} the number of observed $(y_{t-1}, y_{t-2}) \in c_{ij}$. In classifying y_t, we let $x_{ijk} = y_t$ and write the model as $x_{ijk} = f_{ijk} + \varepsilon_{ijk}$ if $(y_{t-1}, y_{t-2}) \in c_{ij}$, where k is used to index the observations in c_{ij}. Under the null hypothesis, $f_{ijk} = f_1(y_{t-1}) + f_2(y_{t-2})$. It is noted first that y_t, y_{t-1}, y_{t-2} have the same marginal stationary distribution and

$$\begin{aligned} v_{ij}(f) &= \operatorname{var}\left[\{f_1(Y_{t-1}) + f_2(Y_{t-2})\} I_{c_{ij}}\right] \\ &\leqslant 2 \max_{i,j} \left[\operatorname{var}\{f_1(Y_t) I_{c_i}\} + \operatorname{var}\{f_2(Y_t) I_{c_j}\}\right] \leqslant 2p_0(L_1^2 + L_2^2)/m^3, \end{aligned}$$

where variance is taken with respect to the stationary density $p(.)$, I_c is the usual indicator function of the set c, $c_i = (a_{i-1}, a_i)$, $L_i = \sup_x \{|f_i'(x)|\}$ and $p_0 = \max_{x \in [a,b]} p(x)$. Hence, under H_0,

$$\frac{1}{n_{ij} - 1} \sum_{k=1}^{n_{ij}} (f_{ijk} - \bar{f}_{ij.})^2 = O_p\{v_{ij}(f)\} = O_p\left(\frac{1}{m^3}\right),$$

where $\bar{f}_{ij.}$ is the average of f_{ijk} over the observations in cell c_{ij}. By expanding the within-sample sum of squares, we obtain that

$$\begin{aligned} \mathrm{SS_W} &= \sum_{i,j} \sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{x}_{ij.})^2 \\ &= \sum_{i,j} \sum_{k} (f_{ijk} - f_{ij.})^2 + \sum_{i,j} \sum_{k} (\varepsilon_{ijk} - \varepsilon_{ij.})^2 + 2 \sum_{i,k} \sum_{k} (f_{ijk} - \bar{f}_{ij.})(\varepsilon_{ijk} - \varepsilon_{ij.}). \end{aligned}$$

Noting that $n_{ij} = O(n^*/m^2)$ and $\sum_k (f_{ijk} - \bar{f}_{ij.})^2 = O(n_{ij}/m^3)$, we have

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n_{ij}} (f_{ijk} - \bar{f}_{ij.})^2 = O_p(n^*/m^3).$$

Furthermore, since the cross-product terms are small and approximately independent with mean zero, we can apply the Linderberg–Feller central limit theorem to obtain

$$(n^* - m^2)^{-\frac{1}{2}} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n_{ij}} (f_{ijk} - \mu_{ij})\varepsilon_{ijk} \to N(0, \sigma^2)$$

in distribution, where $\mu_{ij} = E[I_{c_{ij}}\{f_1(Y_{t-1}) + f_2(Y_{t-2})\}]$ is the true cell mean. Hence the cross-product is of order $O_p\{(n^*/m^3)^{\frac{1}{2}}\}$ because $O(n^* - m^2) = O(n^*)$. Therefore

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{x}_{ij.})^2 = \sum_{i} \sum_{j} \sum_{k} (\varepsilon_{ijk} - \varepsilon_{ij.})^2 + n^* \left[O_p \left\{ \frac{1}{(m^3 n^*)^{\frac{1}{2}}} \right\} + O_p\left(\frac{1}{m^3}\right) \right].$$

Consequently, the larger m is, the better the F approximation. Typically, if m is of order $O(n^{1/3})$ or less

$$\frac{1}{n^* - m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{x}_{ij.})^2 = \frac{1}{n^* - m^2} \sum_{i} \sum_{j} \sum_{k} (\varepsilon_{ijk} - \varepsilon_{ij.})^2 + O_p\left(\frac{1}{m^3}\right).$$

Similar arguments apply to compute the between-sample sum of squares $\mathrm{SS_B}$. Therefore

$$F = \frac{\mathrm{SS_B}/(m-1)^2}{\mathrm{SS_W}/(n^* - m^2)} = F_{(m-1)^2, n^* - m^2} + O_p\left(\frac{1}{m^3}\right),$$

where m is of order $O(n^{1/3})$ or less. This completes the proof. □

Proof of Theorem 2·2. From the previous theorem, it is clear that the local power of the proposed test statistics is the same as that of an ordinary two-way classification with unbalanced cell counts. Let D_1 be the corresponding design matrix for the two-way unbalanced additive analysis of variance model so that D_1 is an $n^* \times (2m-1)$ matrix with rank $2m-1$, and let D_2 be the design matrix for the two-way nonadditive analysis of variance model, so $D_2 = (D_1, D_0)$ is a $n^* \times m^2$ matrix with rank m^2, where D_0 is orthogonal to D_1. Let $X = (x_{ijk})$ be the vector of observations in each cell c_{ij}. Then the residual sums of squares under the two models are

$$S_1^2 = X'\{I - D_1(D_1'D_1)^{-1}D_1'\}X, \quad S_2^2 = X'\{I - D_2(D_2'D_2)^{-1}D_2'\}X.$$

Rewriting (5) as $X = D_1'F + D_0 G m (n^*)^{-\frac{1}{2}} + \varepsilon$, we have $S_2^2 = \varepsilon'\{I - D_2(D_2'D_2)^{-1}D_2'\}\varepsilon$ which is distributed as $\chi^2_{n^* - m^2}$. On the other hand,

$$S_1^2 = \varepsilon'\{I - D_1(D_1'D_1)^{-1}D_1'\}\varepsilon + \frac{m^2}{n^*} \| D_0'G \|^2 + \frac{2m}{(n^*)^{\frac{1}{2}}} \varepsilon' D_0' G,$$

where G is the nonadditive part. Since $n^* = O(n)$ and

$$n^* \min_j (g_j^2) \leqslant \| D_0'G \|^2 \leqslant n^* \max_j (g_j^2),$$

where the g_j's are the coordinates of the vector $D_0'G$, we obtain that $S_1^2 - S_2^2 = \chi^2_{(m-1)^2} + O(m^2)$. The result is proven. □

References

AUESTAD, B. & TJØSTHEIM, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika* **77**, 669–87.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B **36**, 192–236.

BREIMAN, L. & FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Am. Statist. Assoc.* **80**, 580–619.

CHEN, R. & TSAY, R. S. (1993a). Nonlinear additive ARX models. *J. Am. Statist. Assoc.* **88**, 955–67.

CHEN, R. & TSAY, R. S. (1993b). Functional-coefficient autoregressive models. *J. Am. Statist. Assoc.* **88**, 298–308.

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

GRANGER, C. W. J. & ANDERSEN, A. P. (1978). *An Introduction to Bilinear Time Series Models.* Gottingen: Vandenhoek and Ruprecht.

HAGGAN, V. & OZAKI, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* **68**, 189–96.

HASTIE, T. J. & TIBSHIRANI, R. J. (1991). *Generalized Additive Models.* London: Chapman and Hall.

JONES, D. A. (1978). Nonlinear autoregressive processes. *Proc. R. Soc. Lond.* A **360**, 71–95.

KEENAN, D. M. (1985). A Tukey nonadditivity-type test for time series nonlinearity. *Biometrika* **72**, 39–44.

KOYAK, R. (1990). Consistency for ACE-type methods. *Ann. Statist.* **18**, 742–57.

LEWIS, P. A. W. & RAY, B. K. (1993). Nonlinear modeling of multivariate and categorical time series using multivariate adaptive regression splines. In *Nonlinear Time Series and Chaos*, **1**: *Dimension Estimation and Models*, Ed. H. Tong, pp. 136–69. Singapore: World Scientific.

LEWIS, P. A. W. & STEVENS, G. (1991). Nonlinear modelling of time series using multivariate adaptive regression splines (MARS). *J. Am. Statist. Assoc.* **87**, 864–77.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Assoc.* **86**, 316–42.

PRIESTLEY, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis.* London: Academic Press.

ROBINSON, P. M. (1983). Non-parametric estimation for time series models. *J. Time Ser. Anal.* **4**, 185–208.

SUBBA RAO, T. & GABR, M. M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Lecture Notes in Statistics, **24**. New York: Springer-Verlag.

TIBSHIRANI, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *J. Am. Statist. Assoc.* **83**, 394–404.

TJØSTHEIM, D. & AUESTAD, B. (1994). Nonparametric identification of nonlinear time series: Selecting significant lags. *J. Am. Statist. Assoc.* **89**, 1410–9.

TONG, H. (1983). *Threshold Models in Nonlinear Time Series Analysis*, Lecture Notes in Statistics, **21**. Heidelberg: Springer-Verlag.

TONG, H. (1990). *Nonlinear Time Series Analysis: A Dynamical System Approach.* London: Oxford University Press.

TRUONG, Y. K. (1993). A nonparametric framework for time series analysis. In *New Directions in Time Series Analysis*, Ed. D. Billinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt and M. S. Taqqu, pp. 371–86. New York: Springer-Verlag.

TSAY, R. S. (1986). Nonlinearity tests for time series. *Biometrika* **73**, 461–6.

TUKEY, J. W. (1949). One degree of freedom for non-additivity. *Biometrics* **5**, 232–42.

[*Received November* 1993. *Revised July* 1994]