THRESHOLD VARIABLE SELECTION IN OPEN-LOOP THRESHOLD AUTOREGRESSIVE MODELS

By Rong Chen

Texas A&M University

First version received October 1993

Abstract. An open-loop threshold autoregressive model is defined as

$$X_t = \begin{cases} a_0 + \sum_{j=1}^{p} a_j X_{t-j} + \varepsilon_t^{(1)} & \text{if } Z_t < 0 \\ b_0 + \sum_{j=1}^{p} b_j X_{t-j} + \varepsilon_t^{(2)} & \text{if } Z_t \geqslant 0 \end{cases}$$

The main difficulty for building such a model is that the threshold variable Z_t is usually unknown. In practice, there may exist many possible candidates for the threshold variable Z_t. It is difficult and tedious, if not impossible, to search for the best among all the candidates using standard model selection procedures. In this paper, we introduce a digression concept and propose two simple algorithms to classify the observations without knowing the threshold variable. The classification is then used with several graphical procedures to search for the most suitable threshold variable. Simulated and real examples are included to illustrate the proposed procedures.

Keywords. Classification; consistency; digression; nonlinear time series.

1. INTRODUCTION

Many useful nonlinear time series models have been introduced in the literature. Among them, the SETAR models of Tong (1983), the EXPAR model of Haggan and Ozaki (1981), the bilinear model of Subba Rao and Gabr (1984) and the state-dependent model of Priestley (1980) have shown to be capable of modeling various nonlinearities. For detailed information, see Tong (1990) and Priestley (1988). In this paper, we focus on a simplified version of the *open-loop threshold autoregression system* (Tong, 1990, p. 101) which is defined as

$$X_t = \begin{cases} a_0 + \sum_{j=1}^{p} a_j X_{t-j} + \varepsilon_t^{(1)} & \text{if } Z_t < 0 \\ b_0 + \sum_{j=1}^{p} b_j X_{t-j} + \varepsilon_t^{(2)} & \text{if } Z_t \geqslant 0 \end{cases} \tag{1}$$

where $\varepsilon_t^{(i)}$ are white noises with zero mean and finite variances and are independent of Z_t. The variable Z_t, called the *threshold variable*, determines which linear model the process follows at time t. Note that if Z_t is a simple lag variable $X_{t-d} - c$, model (1) reduces to an ordinary SETAR model. Other possibilities of Z_t may be (i) a simple combination of several lag variables, (ii) an exogenous variable or combination of its several lags (e.g. a

0143-9782/95/05 461–481 JOURNAL OF TIME SERIES ANALYSIS Vol. 16, No. 5

leading indicator for modeling an economic series) or (iii) a simple function of a past innovation ε_{t-d}, etc. This extension is more flexible in handling real data sets.

Several methods have been developed for building simple SETAR models. For example, Tsay (1989) selects the possible threshold lag for SETAR models using an F statistic, which was initially designed to test nonlinearity. Tong (1990, p. 405) suggests using profile log likelihood for tentative identification of the threshold variable. These procedures are well designed and easy to use. However, they require the set of possible threshold variables to be specified first and the selection is usually done by considering the candidate threshold variables individually. They are suitable for simple SETAR models since the threshold variable is restricted to a small set of lag variables. But these methods encounter difficulties in dealing with open-loop threshold models. This can be seen from the following example. Suppose the threshold variable is a linear combination of two lag variables, $Z_t = X_{t-1} + \rho X_{t-2} - c$, where ρ is unknown. In order to use the aforementioned methods which require ρ to be given, one has to try out every possible ρ. This can be very tedious, if not impossible. The main objective of this paper, therefore, is to propose simple yet efficient procedures to overcome the difficulty.

The advantages of knowing the threshold variable is that all the observations can then be easily classified into different regimes. That is, the index set $\{1, \ldots, n\}$ can be partitioned into two sets T_1 and T_2, based on the known threshold variable. Within each set, the observations follow a single autoregressive model. In this paper we propose a method that reverses the above approach. First, without knowing or assuming the threshold variable, we obtain the partition T_1 and T_2 from the data. This is done by fitting two regression surfaces simultaneously to the data and is referred to as *digression* (Kotz *et al.*, 1988, Vol. 2, p. 373). Then we try to find a threshold variable which is suitable for the given partition. Because the partition is fixed, the search becomes much easier and does not involve model building and comparison procedures. With graphics, it is even possible to find a simple function such as $Z_t = X_{t-1} - \rho X_{t-2}^2$ as the threshold variable.

The rest of the paper is organized as follows. In Section 2, the concept of digression is introduced and two classification algorithms are proposed with simulation studies. A consistency result is also given. In Section 3, several procedures are proposed for searching for the threshold variable. Simulated examples are used to illustrate the procedures. A real data example is given in Section 4. The proofs of the theorems are given in the appendix.

2. CLASSIFICATION VIA DIGRESSION

2.1. *Digression*

The basic setup of digression is as follows. Assume $(Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, following

$$Y_i = \begin{cases} a_0 + \sum_{j=1}^{p} a_j X_{ij} + \varepsilon_i^{(1)} & \text{if } I_i = 1 \\ b_0 + \sum_{j=1}^{p} b_j X_{ij} + \varepsilon_i^{(2)} & \text{if } I_i = 2. \end{cases} \tag{2}$$

where $\varepsilon_i^{(k)}$ i.i.d. $\sim \mathrm{N}(0, \sigma_k^2)$, $k = 1, 2$. The indicator variables I_i are latent variables and are assumed to be independent of the noise $\varepsilon_i^{(k)}$. In addition, we assume $P(I_i = 1) = \theta_i$, $i = 1, \ldots, n$, where θ_i are nonstochastic unknown parameters. The model setting here is different from general switching regression models. For example, Quandt and Ramsey (1978) and Shumway and Stoffer (1991) assume that the I_is are independent and identically distributed (i.i.d.) Bernoulli random variables with $P(I_i = 1) = \theta$. Goldfeld and Quandt (1973) assume that the I_is follow a first-order two-state Markov chain. Here we do not have any assumptions on the parameters $\theta_1, \ldots, \theta_n$.

If each observation is forced to be attributed to the nearest regression surface, i.e. $\theta_i = 0$ or 1, then, with observations $\{y_i, x_{i1}, \ldots, x_{ip}\}$, $i = 1, \ldots, n$, the *selective least squares* estimators are the a_j, b_j, $i = 0, \ldots, p$, that minimize

$$S(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} \min\left\{\left(y_i - a_0 - \sum_{j=1}^{p} a_j x_{ij}\right)^2, \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2\right\}. \tag{3}$$

It is easy to prove that (3) is equivalent to

$$S(\boldsymbol{a}, \boldsymbol{b}) = \min_{T_1, T_2}\left\{\sum_{i \in T_1}\left(y_i - a_0 - \sum_{j=1}^{p} a_j x_{ij}\right)^2 + \sum_{i \in T_2}\left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2\right\} \tag{4}$$

where the minimization takes over all possible partitions T_1 and T_2 of the index set $\{1, \ldots, n\}$. Note that neither (3) nor (4) are easy to minimize since there are 2^{n-1} ways to partition the index set $\{1, \ldots, n\}$, after eliminating the symmetric situations. In addition, the selective least squares estimator is not consistent. The inconsistency is due to the fact that, as the sample size grows, the portion of misclassifications does not go to zero, neither does the bias introduced by them. The proof is given in the appendix.

2.2. *Classification algorithm I – discarding algorithm*

In this section, we propose a classification algorithm which provides consistent estimators. To illustrate the proposed algorithm, we accompany the procedure with an example shown in Figure 1, which is a scatterplot of observations from two regression lines. The proposed algorithm can be formulated as follows.

(1) Find good initial parameter values through the following procedure.

(i) Partition the data range of each explanatory variable into k equal intervals. By doing this, we obtain k^p blocks in the p-dimensional explanatory space. The choice of k depends on the sample size and the dimension p. Large k is preferred, but it has to be controlled so that there are a fair

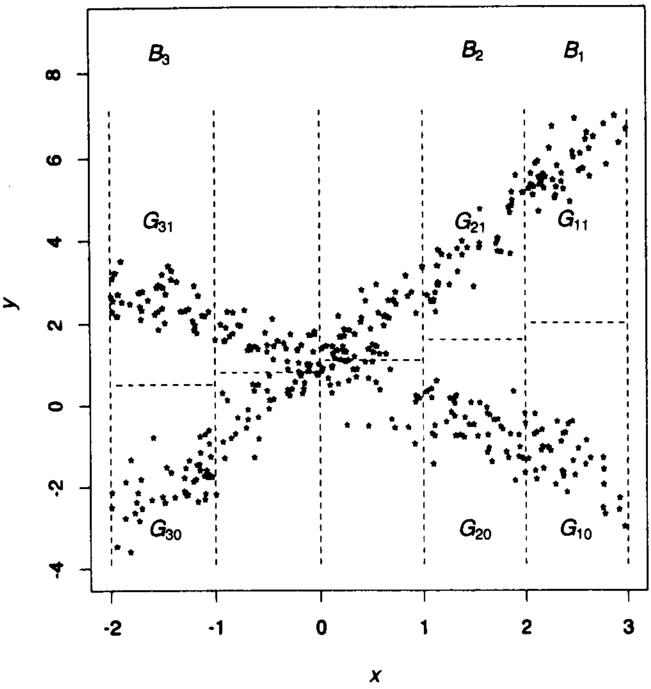


FIGURE 1. Illustration of Step 1 in the discarding algorithm.

number (about $3p$ or more) of blocks having reasonable sample sizes (say, more than 30 observations). The blocks with very few observations should be excluded from further consideration. In Figure 1, we partition the one-dimensional explanatory space into five blocks.

(ii) Compute the variances of response variable Y in each block and sort them in decreasing order. Let the first l blocks be $\{B_1, \ldots, B_l\}$. The number l should be large enough to estimate a p-dimensional surface, but not too large in order to eliminate the blocks with relatively small variances. In addition, since the amount of computation in the following steps grows exponentially with l, large l should be avoided. Experience shows that l between $p+2$ and $2p$ may be enough. For the data set in Figure 1, we choose $l=3$. The first three blocks, in the order of sample variances, are marked in the figure.

(iii) Within each block B_j, $j=1, \ldots, l$, separate the observations into two groups G_{j0} and G_{j1}, such that the sum $s^2(G_{j0})+s^2(G_{j1})$ is minimum over all the possible separations, where $s^2(G_{j0})$ and $s^2(G_{j1})$ are sample variances of the response variable Y in the two groups. In Figure 1, the separations are shown by the broken lines.

(iv) Select one group from each of the l selected blocks, say, $\{G_{1i_1}, \ldots, G_{li_l}\}$, where i_k is either 0 or 1. Use all the observations in those l groups to estimate the first regression surface. The remaining l groups $\{G_{1(1-i_1)}, \ldots, G_{l(1-i_l)}\}$ are used to estimate the second regression surface. Compute the overall residual sum of squares of the two regressions. Note that this depends on the selection of $\{i_1, \ldots, i_l\}$. We denote it $S(i_1, \ldots, i_l)$. Due to symmetry, we assume $i_1 = 1$ without loss of generality. Compute $S(i_1, \ldots, i_l)$ for all 2^{l-1} possible combinations. Estimates corresponding to the minimum S among those possible combinations are used as the initial estimates. In other words, the initial estimates $\tilde{a}_k$, $\tilde{b}_k$, $k = 0, \ldots, p$, are those that minimize

$$\min_{(i_1,\ldots,i_l)} \sum_{j=1}^{l} \left\{ \sum_{(y,X)\in G_{ji_j}} \left(y - a_0 - \sum_{k=1}^{p} a_k x_k\right)^2 + \sum_{(y,X)\in G_{j(1-i_j)}} \left(y - b_0 - \sum_{k=1}^{p} b_k x_k\right)^2 \right\}.$$

In Figure 1, $S(1,1,0)$ is the minimum among the four possible combinations $S(1,0,0)$, $S(1,0,1)$, $S(1,1,0)$ and $S(1,1,1)$. Hence, the initial estimates of the first regression line are the ordinary least squares regression estimates using *only* the observations in groups G_{11}, G_{21} and G_{30} while the initial estimates of the second regression are those using the observations in groups G_{10}, G_{20} and G_{31}.

(2) Find consistent estimates.

(i) Select a constant M and the error margin $(v_0, \ldots, v_p)$ of the initial estimates, i.e. we assume $|\tilde{a}_i - a_i| < v_i$, $|\tilde{b}_i - b_i| < v_i$, $i = 0, \ldots, p$, where $\tilde{a}_i$, $\tilde{b}_i$ are the initial estimates and a_i, b_i are the underlying true coefficients. Some guidelines for choosing those parameters are given in the remarks below.

(ii) For each observation $\{y_j, x_{j1}, \ldots, x_{jp}\}$, compute the fitted values $r_{j1} = \tilde{a}_0 + \sum_{i=1}^{p} \tilde{a}_i x_{ji}$ and $r_{j2} = \tilde{b}_0 + \sum_{i=1}^{p} \tilde{b}_i x_{ji}$ and the corresponding residuals $e_{j2} = y_j - r_{j2}$. For each observation, also compute $E_j = 2(\sum_{i=1}^{p} v_i |x_{ji}| + v_0)$ which measures the maximum distance between the initially estimated regression surfaces to the 'true' ones on point $(x_{j1}, \ldots, x_{jp})$.

(iii) Let

$$T_1 = \{j\colon e_{j1}^2 < e_{j2}^2, |r_{j1} - r_{j2}| > M + E_j\}$$

and

$$T_2 = \{j\colon e_{j2}^2 < e_{j1}^2, |r_{j1} - r_{j2}| > M + E_j\}.$$

Then the ordinary linear regression estimates $\hat{a}_i$, $i = 0, \ldots, p$, and $\hat{b}_i$, $i = 0, \ldots, p$, obtained from the sets T_1 and T_2, respectively, are regarded as the estimators of the coefficients a_i and b_i, $i = 0, \ldots, p$.

Several remarks are in order.

(i) The final estimators are consistent, given reasonable initial values. The detailed conditions are given later in the appendix. The key of the algorithm is that the probability of misclassification decreases to zero as M goes to infinity. Hence, as the sample size increases, if we allow the constant M to increase to reduce the bias due to the misclassifications, we can achieve consistency. However, the procedure is not efficient since it discards a portion of the observations.

(ii) The results hold under the conditions that the sequence (X_i, I_i) is a finite-order Markov chain and is stationary and ergodic. Hence I_i can be a function of some lag variables of X_i or other stationary exogenous variables. Stationary and ergodic threshold models belong to this class, as well as the open-loop threshold models.

(iii) The constant M reflects the compromise between the consistency and the efficiency, or the bias and the variance. The optimal selection of M depends on the sample size, the distribution of X and the initial values and can be obtained by minimizing the mean squared error

$$\frac{\{\mathrm{tr}\,(\hat{\Sigma}_1^{-1}[\Sigma_{i\in T_1(M)} X_i' X_i \Phi\{-|X_i(\alpha_0 - \beta_0)|/\hat{\sigma}_2\}])\}^2 \|\alpha_0 - \beta_0\|^2}{n_1^2} + \frac{\hat{\sigma}_1^2 \,\mathrm{tr}\,(\hat{\Sigma}_1^{-1})}{n_1}$$
$$+ \frac{\{\mathrm{tr}\,(\hat{\Sigma}_2^{-1}[\Sigma_{i\in T_2(M)} X_i' X_i \Phi\{-|X_i(\alpha_0 - \beta_0)|/\hat{\sigma}_1\}])\}^2 \|\alpha_0 - \beta_0\|^2}{n_2^2} + \frac{\hat{\sigma}_2^2 \,\mathrm{tr}\,(\hat{\Sigma}_2^{-1})}{n_2}$$

where $\hat{\Sigma}_i$, $\hat{\sigma}_i^2$ and n_i are the sample covariance matrix of the explanatory variables, the residual sample variance and sample size in $T_i(M)$, respectively. The derivation is easy and hence omitted. By assuming that the covariance matrices $\hat{\Sigma}_i$ are the same as the covariance matrix $\hat{\Sigma}$ using all the observations and that $\hat{\sigma}_i^2$ does not change much with M, we can obtain an approximation of the above expression:

$$\left[\left\{\Phi\left(-\frac{M}{2\hat{\sigma}_2}\right)\right\}^2 + \left\{\Phi\left(-\frac{M}{2\hat{\sigma}_1}\right)\right\}^2\right]\|\hat{\alpha}_0 - \hat{\beta}_0\|^2 + \left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)\mathrm{tr}\,(\hat{\Sigma}^{-1}).$$

Since larger M results in smaller n_1 and n_2, the competition between the two terms can been seen immediately.

(iv) The error margins υ_i express how strong one believes in the initial estimates. Small error margins υ_i increase the efficiency of the estimator since fewer observations are discarded. The standard errors of the initial estimates can be used as a rough guideline of the error margins.

(v) The intuition behind Step 1 is that it is easier and more accurate to separate the observations from the two models in the blocks with large variance in the response variable. In those blocks, the distances between the observations from the two models are relatively large. Due to the 'curse of dimensionality', a large sample size is needed to perform Step 1 when the dimension of the covariates p is large. However, it is possible to obtain correct information of the classification when only a small number of most significant covariates are used in the procedure.

2.3. *Classification algorithm II – Bayesian algorithm*

In this section we introduce another classification algorithm from a Bayesian viewpoint. For simplicity of notation, we rewrite model (2) as

$$y_i = \begin{cases} \alpha' X_i + \varepsilon_i^{(1)} & \text{if } I_i = 1 \\ \beta' X_i + \varepsilon_i^{(2)} & \text{if } I_i = 2 \end{cases}$$

where α and β are the $(p + 1)$-dimensional coefficient vectors, including the constant term and $X_i' = (1, x_{i1}, \ldots, x_{ip})$. Denote the set of all observations by $\mathcal{Y}$. If we assume flat prior distributions on the regression parameters α, β and the classification parameters $\theta_i = P(I_i = 1)$, then the posterior distribution of the indicators I_i is, after integrating out the parameters α, β, σ_k^2 and the θ_i's,

$$\pi(I_1, \ldots, I_n|\mathcal{Y}) \propto$$

$$\frac{\Gamma\{(n_1 - p - 3)/2\}\Gamma\{(n - n_1 - p - 3)/2\}}{(|S_{xx1}S_{xx2}|)^{1/2}} \Delta_1^{(n_1-p-1)/2} \Delta_2^{(n-n_1-p-1)/2} \tag{5}$$

where $n_1 = \sum_{i=1}^{n}(2 - I_i)$ is the number of observations such that $I_i = 1$, and

$$S_{xx1} = \sum_{i=1}^{n}(2 - I_i)X_iX_i' \qquad S_{xy1} = \sum_{i=1}^{n}(2 - I_i)X_iy_i$$

$$\Delta_1 = \sum_{i=1}^{n}(2 - I_i)y_i^2 - (S_{xy1})'(S_{xx1})^{-1}(S_{xy1})$$

$$S_{xx2} = \sum_{i=1}^{n}(I_i - 1)X_iX_i' \qquad S_{xy2} = \sum_{i=1}^{n}(I_i - 1)X_iy_i$$

$$\Delta_2 = \sum_{i=1}^{n}(I_i - 1)y_i^2 - (S_{xy2})'(S_{xx2})^{-1}(S_{xy2}).$$

The derivation of the above formula follows from a straightforward integration procedure using properties of normal distributions. To make inference using (5), we employ the Gibbs sampler to obtain random draws from the posterior distribution. The Gibbs sampler is a powerful Markov chain Monte Carlo technique which has been widely used recently. For detailed information, see Gelfand and Smith (1990). The Gibbs sampler is implemented as follows. For a fixed i, the conditional probability $\pi(I_i = 1|I_1, \ldots, I_{i-1}, I_{i+1}, \ldots, I_n, \mathcal{Y})$ can be computed easily using (5). Then a Bernoulli random variate is drawn according to this probability and I_i is updated accordingly. Starting with trivial values, the procedure runs from $i = 1$ to n to complete one iteration of the Gibbs sampler. In practice, $N_1 + N_2$ iterations are needed, among which the first N_1 iterations are discarded and the last N_2 iterations are saved as the draws from the posterior distribution (5). Since the iterative procedure 'walks' on a discrete set of 2^n points, geometric convergence is guaranteed (Liu *et al.*, 1995). In addition, the computation is easy in the above procedure since iterative formulas for matrix inversions and determinant computation are available. For details, see Chen and Liu (1995).

To obtain the coefficient estimates, one can either adopt an empirical Bayes approach by obtaining the least squares estimates corresponding to the posterior mode, or use the observations with posterior probability $P(I_i = 1) > p$ to estimate one regression surface and those with $P(I_i = 1) < 1 - p$ to estimate the other, where p is a predetermined constant. Intuitively the choice of p is again a compromise between the bias and the variance. Large p reduces the bias but increases the variance since fewer observations are used.

2.4. *Simulated examples*

A simulation study is carried out to explore the properties of the proposed procedures. The following simple open-loop threshold AR(2) model is used.

$$x_t = \begin{cases} 0.7x_{t-1} - 0.3x_{t-2} + \varepsilon_t & \text{if } z_t < 0 \\ 1 - 0.7x_{t-1} - 0.8x_{t-2} + \varepsilon_t & \text{if } z_t \geqslant 0 \end{cases}$$

with four different threshold variables: I, z_t i.i.d. $\sim \mathrm{N}(0, 1)$; II, $z_t = x_{t-2}$; III, $z_t = x_{t-3} + x_{t-4}^2$; and IV, $z_t = x_{t-2} + x_{t-3} + x_{t-4}$. With each model, we generated 500 series, each with 400 observations. The noise ε_t is generated from $\mathrm{N}(0, 0.5^2)$. Then the proposed discarding algorithm and the Bayesian algorithm are used to estimate the coefficients in the model. For comparison reasons, we also computed the least squares estimates of the coefficients using the 'true' threshold variable. The initial estimates are obtained by partitioning the two-dimensional explanatory variable space $\{x_{t-1}, x_{t-2}\}$ into 25 blocks and using the four blocks with the largest variances (Step 1 of the discarding algorithm.) The selection of those method parameters is based on the suggestions in Section 2.2. The final estimates are the results of Step 2 using constant $M = 1$ and error margins $v_i = 0.1$. The Bayesian estimates are obtained by running the Gibbs sampler 1000 iterations. The first 500 iterations are discarded and the rest are used to compute the posterior probability $P(I_i = 1)$. The coefficient estimates are obtained using the observations with $P(I_i = 1) > 0.95$ and $P(I_i = 1) < 0.05$.

Table I shows the mean and standard deviation (in parentheses) of the estimated coefficients of the 500 simulated series for each model. We can see that the estimates we obtained (both through the discarding algorithm and the Bayesian algorithm), without any knowledge of the threshold variable, are reasonably close to the estimates under the exact true threshold variable, though the sampling variation is slightly higher.

Discussion. As we have seen, the Bayesian algorithm is an automatic algorithm. It does not require any 'good' initial parameter values. However, it is computationally expensive. On the other hand, the discarding algorithm depends heavily on the initial parameter values. There are many situations in which Step 1 of the discarding algorithm fails to provide reasonable initial values, particularly when the sample size is small and the observations of the explanatory variables are irregularly spaced.

TABLE I

SIMULATION RESULTS

Model	'True' estimates	Initial estimates	Final estimates	Bayesian estimates
I	1.00, −0.70, −0.80	1.02, −0.70, −0.76	1.01, −0.70, −0.79	1.03, −0.73, −0.80
	(0.04, 0.03, 0.03)	(0.13, 0.11, 0.09)	(0.07, 0.04, 0.04)	(0.10, 0.10, 0.06)
	0.00, 0.70, −0.30	0.00, 0.68, −0.29	0.00, 0.69, −0.30	−0.02, 0.73, −0.30
	(0.04, 0.03, 0.03)	(0.12, 0.11, 0.09)	(0.07, 0.04, 0.05)	(0.11, 0.10, 0.07)
II	1.00, 0.70, −0.80	1.07, −0.64, −0.63	1.04, −0.66, −0.76	1.06, −0.75, −0.71
	(0.06, 0.03, 0.10)	(0.22, 0.19, 0.29)	(0.21, 0.22, 0.18)	(0.12, 0.06, 0.13)
	0.00, 0.69, −0.30	−0.02, 0.46, −0.30	−0.02, 0.58, −0.38	−0.18. 0.84, −0.20
	(0.05, 0.05, 0.04)	(0.25, 0.35, 0.24)	(0.26, 0.26, 0.36)	(0.12, 0.11, 0.12)
III	1.00, −0.70, −0.80	0.96, −0.60, −0.82	0.94, −0.67, −0.85	1.03, −0.75, −0.83
	(0.05, 0.04, 0.05)	(0.19, 0.21, 0.21)	(0.12, 0.12, 0.13)	(0.09, 0.05, 0.09)
	0.00, 0.70, −0.30	−0.07, 0.63, −0.26	−0.04, 0.68, −0.30	−0.09, 0.77, −0.31
	(0.03, 0.03, 0.03)	(0.11, 0.15, 0.11)	(0.07, 0.07, 0.06)	(0.06, 0.05, 0.05)
IV	1.00, −0.70, −0.80	1.01, −0.68, −0.76	0.96, −0.69, −0.84	0.96, −0.75, −0.86
	(0.06, 0.03, 0.06)	(0.21, 0.15, 0.20)	(0.13, 0.08, 0.12)	(0.16, 0.04, 0.14)
	0.00, 0.69, −0.30	−0.10, 0.53, −0.26	−0.08, 0.62, −0.29	−0.10, 0.70, −0.29
	(0.03, 0.04, 0.03)	(0.16, 0.23, 0.12)	(0.10, 0.12, 0.08)	(0.15, 0.11, 0.08)

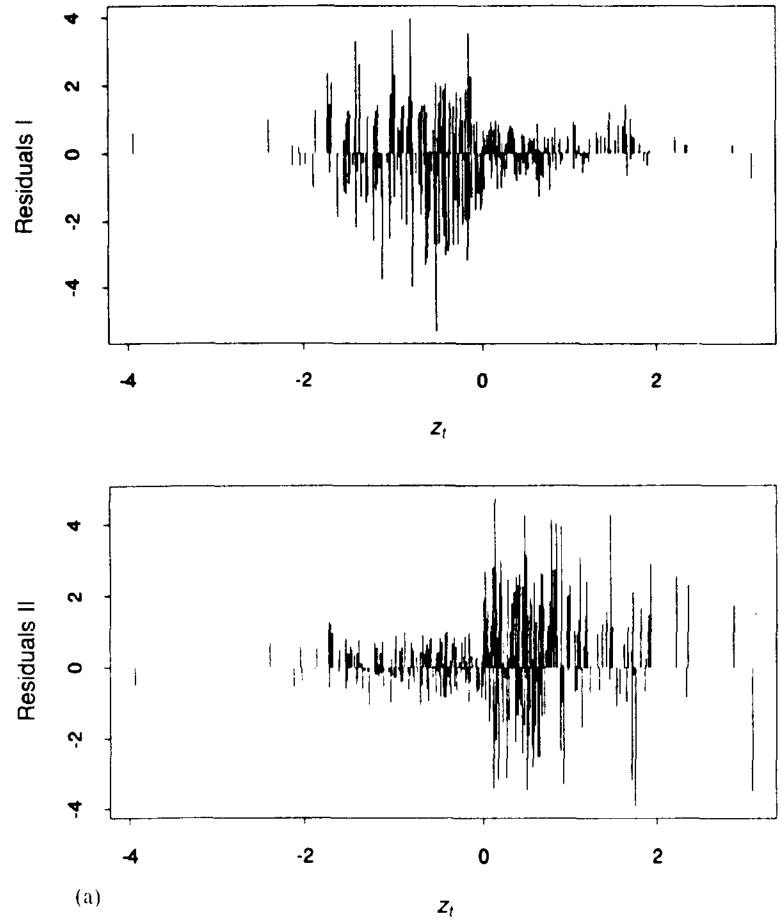
3. SEARCHING FOR THRESHOLD VARIABLES

After the index set $\{1, \ldots, n\}$ is partitioned using the proposed algorithms, a searching procedure can be applied to find a suitable threshold variable either graphically or numerically. Rough comparisons of different threshold variables can be made easily.

3.1. *Graphical tools using residuals and posterior probabilities*

For each observation, we can compute the residuals from the two regression surfaces obtained through the classification algorithms. (We shall use e_{j1} and e_{j2} to denote the residuals throughout the remainder of this paper.) Then the residuals can be plotted against the candidates of threshold variable. Figure 2 is such a residual plot using different candidate threshold variables for one realization of Model I in Section 2.4. Figure 2(a) uses the true threshold variable z_t while Figure 2(b) uses an 'incorrect' threshold variable z_{t-1}. Figure 2(a) clearly indicates that x_t belongs to one regression surface for $z_t < c$ and to another for $z_t > c$. (c is close to 0 from the graph.) Hence, x_t follows a threshold model with z_t as the threshold variable. Meanwhile, 'large' residuals and 'small' residuals are mixed in Figure 2(b) which shows that z_{t-1} is an incorrect threshold variable. Figure 3 plots the posterior probability of $P(I_i = 1)$ against the candidate threshold variables. We note that, when the posterior probability is plotted against the corrected threshold variables, there are empty spaces in the upper left and lower right corners.

If the threshold variable is suspected to be a simple combination of two



(a)

variables z_1 and z_2, then the following simple procedure can be used. First, compute the residuals e_{i1} and e_{i2}. Define

$$d_i = \begin{cases} + & \text{if } |e_{i1}| > |e_{i2}| + \delta \\ \cdot & \text{if } -\delta < |e_{i1}| - |e_{i2}| < \delta \\ - & \text{if } |e_{i2}| > |e_{i1}| + \delta \end{cases} \tag{6}$$

and plot (z_1, z_2) using d_i as characters. The '·' indicates that those observations are not important when we try to separate the observations and the δ can be adjusted. Figure 4 shows such a graph using a realization of Model III in Section 2.4. The figure uses (x_{t-3}, x_{t-4}) and $\delta = 0.2$. It is clear that a quadratic function separates the '+' and '−' signs quite well although the number of observations in the '−' region is small. Figure 5 plots the residuals against the true threshold variable $x_{t-3} + x_{t-4}^2$. It shows a threshold of 0. Similar plots can be generated using the posterior probabilities.

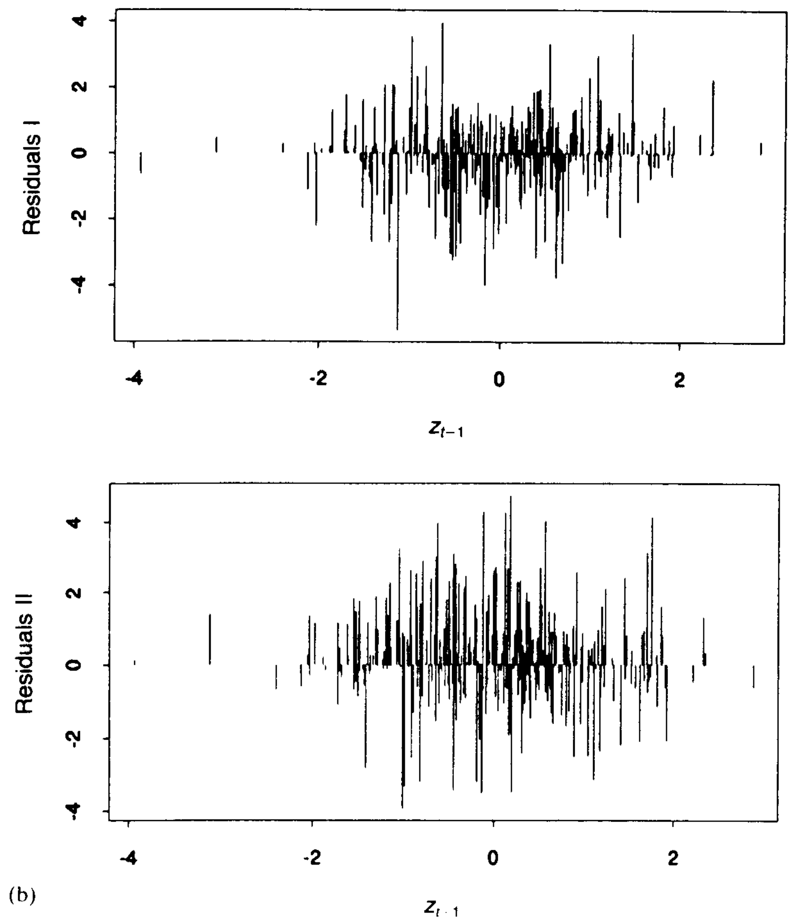


FIGURE 2. (a) Residuals from two regression surfaces against the true threshold variable z_t for one realization of simulated Model I; (b) residuals from two regression surfaces against incorrect threshold variable z_{t-1} for one realization of simulated Model I.

3.2. *Numerical comparison*

The sum of squares of residuals can be used to select 'best' threshold variables given a particular classification. If the candidate threshold variable under study is $Z - c$, where c is an unknown threshold value, define

$$r_z^2 = \min_c \sum_{i=1}^{N} [e_{i1}^2 \delta_c(z_i) + e_{i2}^2\{1 - \delta_c(z_i)\}] \qquad (7)$$

where $\delta_c(z) = 1$ if $z < c$ and zero otherwise. Since the residuals e_{i1} and e_{i2} are fixed, the computation of r_z^2 is easy for a given Z. If Z is a linear combination of two or three variables, the coefficients of the combination can be searched at reasonable speed if graphical tools can provide narrow bounds for these coefficients. In summary, after the classifications are obtained

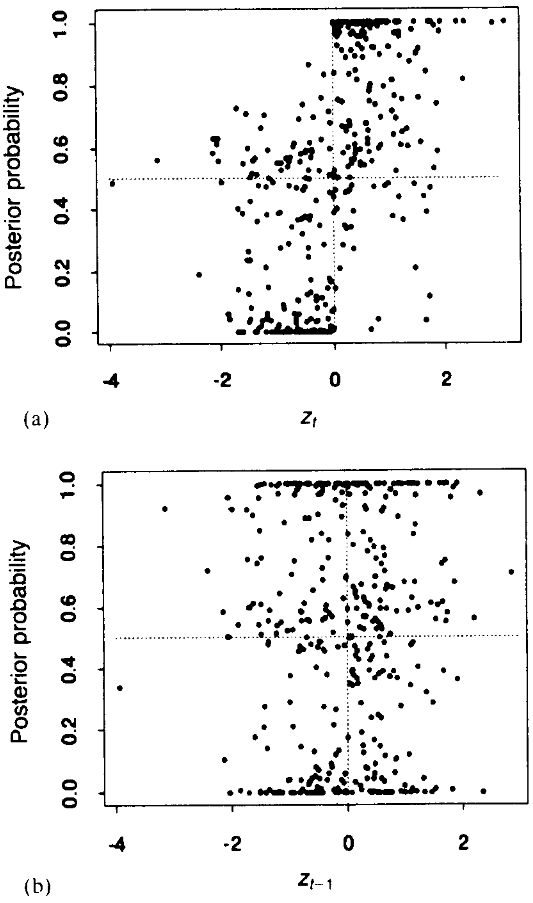


FIGURE 3. (a) Posterior probabilities $P(I_t = 1)$ against the true threshold variable z_t for one realization of simulated Model I; (b) posterior probabilities $P(I_t = 1)$ against the incorrect threshold variable z_{t-1} for one realization of simulated Model I.

through the digression procedure, the search for the threshold variable becomes an easy task.

4. A REAL EXAMPLE

In this section we apply the proposed procedure to the yearly Wolf sunspot number from year 1700 to 1978. A transformation of $2\{(1 + x_t)^{1/2} - 1\}$ is used, following Ghaddar and Tong (1981). This data set has been analyzed extensively in the nonlinear time series literature. See Ghaddar and Tong (1981), Izenman (1983), Subba Rao and Gabr (1984) and Lewis and Stevens (1991). Here we use it as an illustration of the proposed procedures. Using two explanatory variables x_{t-i_1} and x_{t-i_2} and searching over all possible

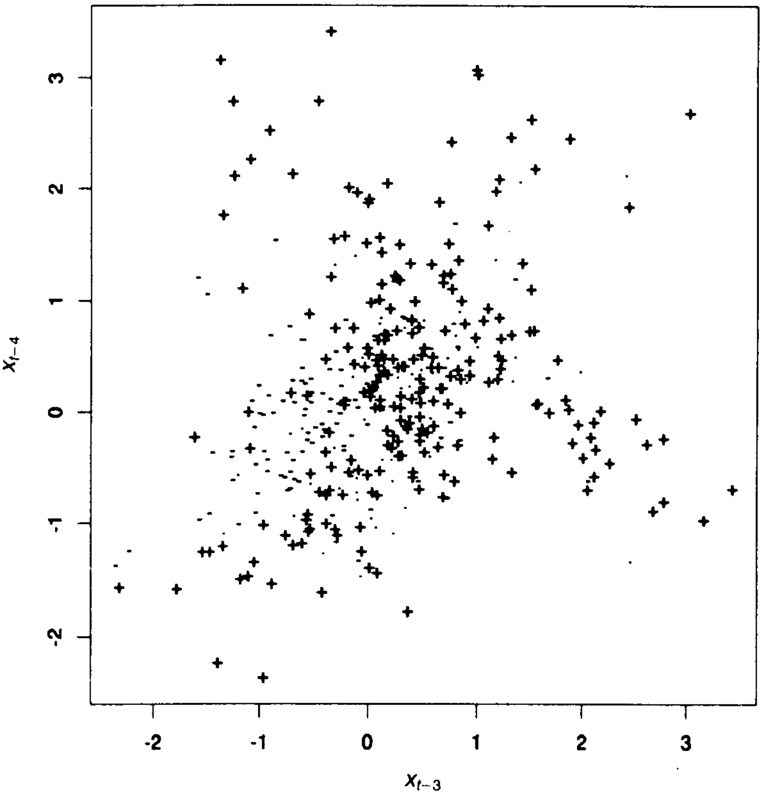


FIGURE 4. Plot of (x_{t-3}, x_{t-4}) for one realization of simulated model III: +, residual from the first regression surface is smaller than that from the second; −, otherwise.

combinations of i_1 and i_2 $(i_1, i_2 \leq 10)$, where the initial estimates are obtained using four blocks with largest variances in 16 partitioned blocks for all lag combinations, the discarding algorithm finds that $i_1 = 1$ and $i_2 = 8$ provide the smallest residual sum of squares with final estimates using $M = 1$ and $v = 0.5$. The sum of squares of residuals is computed by $\sum_{j=1}^{n} \min(e_{j1}^2, e_{j2}^2)$. The final estimates are

$$\text{Model 1} \quad x_t = -2.533 + 0.978x_{t-1} + 0.091x_{t-8}$$

$$\text{Model 2} \quad x_t = 1.143 + 0.804x_{t-1} + 0.304x_{t-8}$$

with residual sum of squares 691.98.

Using the graphical tools described in Section 3.1, we checked all the single lag x_{t-d} and combinations of two lags x_{t-d_1} and x_{t-d_2} for possible threshold variables. A combination of x_{t-1} and x_{t-3} (Figure 6) provides the most clearly separated '+' and '−' signs where a quadratic function is suggested. Figure 7 also confirms the finding. In Figure 7, '+' is used to denote an observation with posterior probability $P(I_i = 1) > 0.6$, '−' an observation with posterior probability $P(I_i = 1) < 0.4$ and '·' all other observations. The posterior

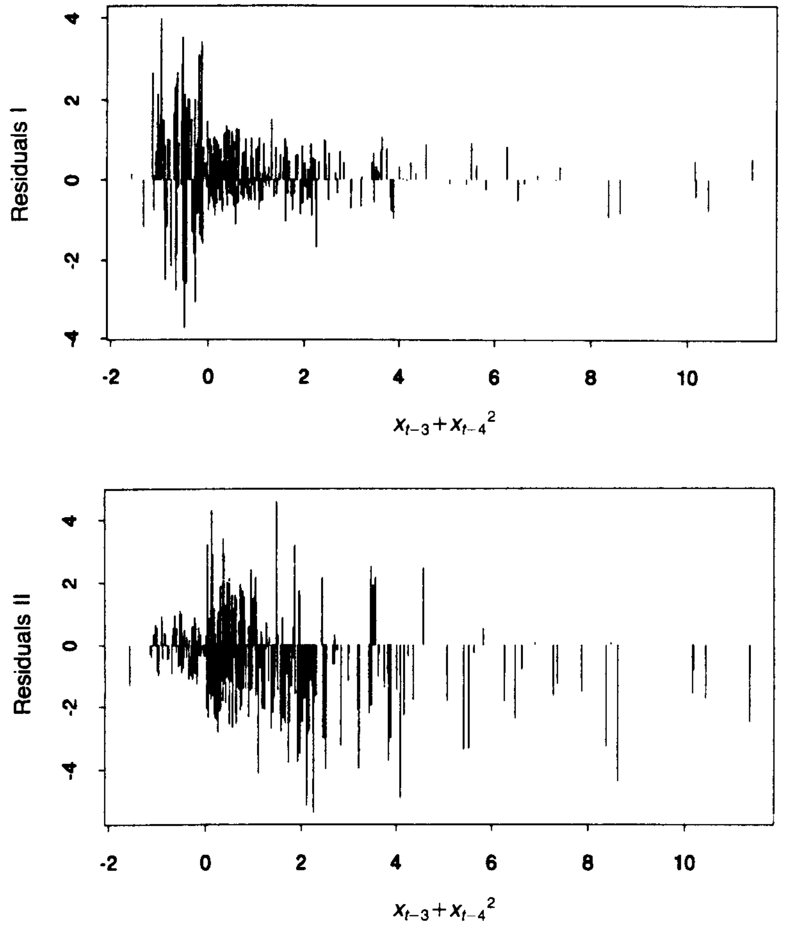


FIGURE 5. Residuals from two regression surfaces against the true threshold variable $z_t = x_{t-3} + x_{t-4}^2$ for one realization of simulated Model III.

probabilities are obtained by running the Gibbs sampler 6000 iterations with the first 5000 iterations discarded.

We then searched the coefficient of this possible quadratic function and found that

$$z_t = 10.25x_{t-3} + (x_{t-1} - 10.25)^2 - 111.86 \tag{8}$$

provides the smallest r_z^2 as defined in (7).

Since the sample size in this example is relatively small, our procedures cannot reliably handle the cases with more than two lag variables, due to the 'curse of dimensionality'. But a model with only two lag variables in the autoregressive equation is hardly satisfactory for a complex process like the sunspot numbers. In order to improve it, we adopt an *ad hoc* procedure. First

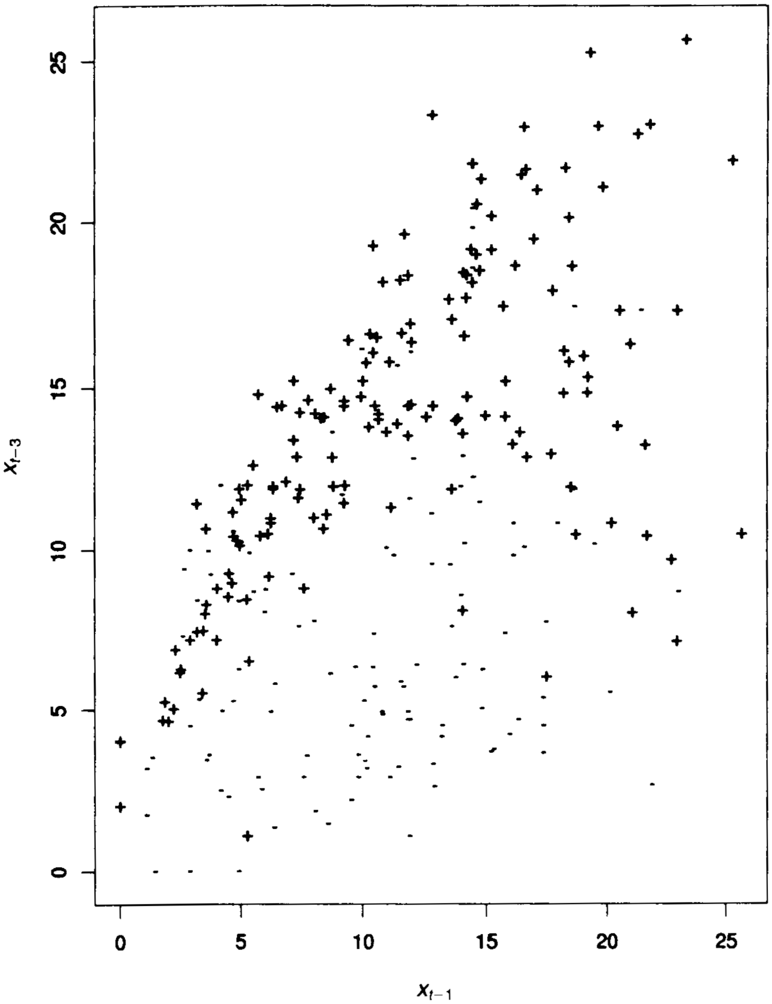


FIGURE 6. Plot of (x_{t-1}, x_{t-3}) for the sunspot example: +; residual from the first regression surface is smaller than that from the second; −, otherwise.

ve fix the threshold variable as that in (8) and add several lags to both utoregressive equations according to the Akaike information criterion AIC). The final model we obtained is

$$x_t = \begin{cases} a_0 + a_1x_{t-1} + a_2x_{t-2} + a_3x_{t-3} + a_4x_{t-4} + a_8x_{t-8} + \varepsilon_t^{(1)} & \text{if } z_t < 0 \\ b_0 + b_1x_{t-1} + b_2x_{t-2} + b_7x_{t-7} + b_8x_{t-8} + b_9x_{t-9} + \varepsilon_t^{(2)} & \text{if } z_t \geqslant 0 \end{cases} \tag{9}$$

here $z_t = (x_{t-1} - c_1)^2 + c_2x_{t-3} + c_3$. After all the variables are fixed in the nodel, we reestimated all the parameters, including the structure parameters the threshold variable. The conditional least squares estimates and their

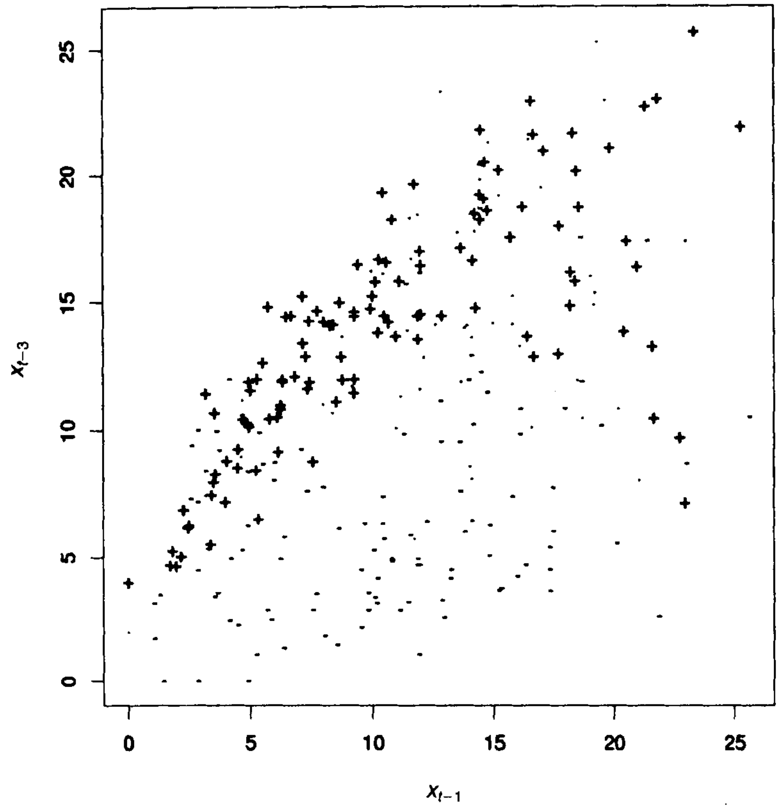


FIGURE 7. Plot of (x_{t-1}, x_{t-3}) for the sunspot example: +, posterior probability $P(I_t = 1) < 0.4$; –, $P(I_t = 1) > 0.6$; ·, otherwise.

standard errors (in parentheses) are

$$\begin{array}{lll}
\hat{a}_0 = 0.490\ (0.866) & \hat{a}_1 = 1.453\ (0.098) & \hat{a}_2 = -0.790\ (0.165) \\
\hat{a}_3 = 0.300\ (0.176) & \hat{a}_4 = -0.150\ (0.125) & \hat{a}_8 = 0.217\ (0.056) \\
\hat{b}_0 = 0.133\ (0.649) & \hat{b}_1 = 1.010\ (0.060) & \hat{b}_2 = -0.255\ (0.068) \\
\hat{b}_7 = 0.036\ (0.061) & \hat{b}_8 = -0.158\ (0.097) & \hat{b}_9 = 0.295\ (0.066)
\end{array}$$

$$\hat{c}_1 = 10 \qquad \hat{c}_2 = 10 \qquad \hat{c}_3 = -113$$

The structure parameter estimates $\hat{c}_1$, $\hat{c}_1$ and $\hat{c}_3$ are obtained by minimizing the overall residual sum of squares. Figure 8 shows the fitted plot and the residual plot. The residual variances for the two regimes are 4.13 with 89 observations and 3.42 with 182 observations, respectively. The overall residual variance is 3.64.

It is difficult to interpret the strange threshold variable in (8). Figure 9 plots the sunspot series, where the observations in the first regime are circled. We can see that most of the circled observations are in the late part of the

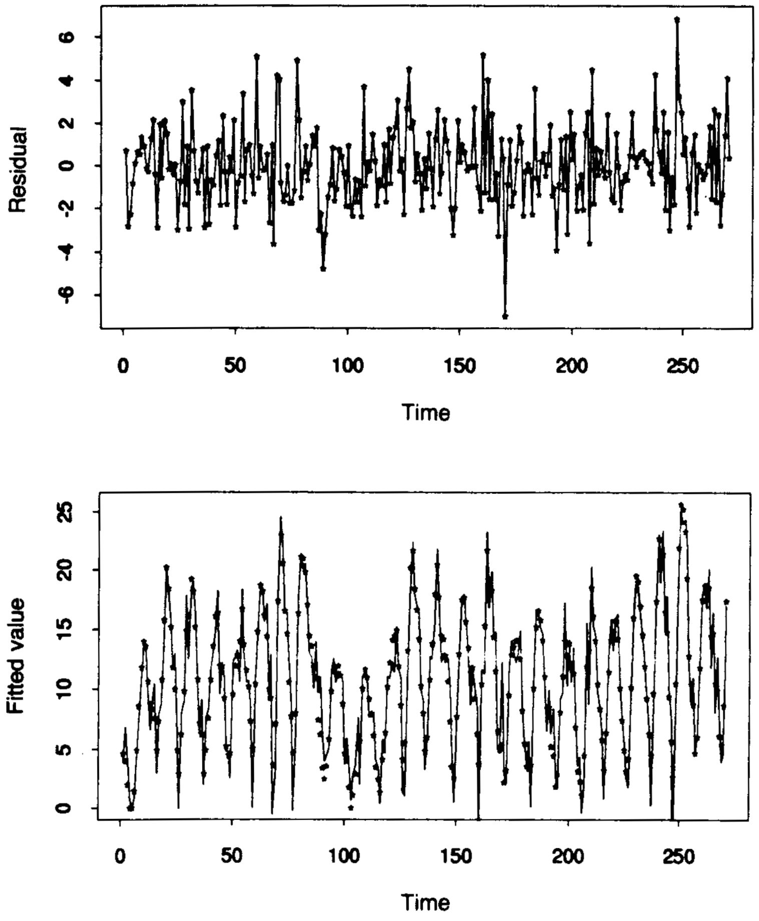


FIGURE 8. Residual plot and fitted value plot for the sunspot example.

ascent period. Peaks usually belong to the first regime while the valleys do not.

APPENDIX

Here we prove that the discarding algorithm provides consistent estimates of the coefficients in (2) and argue that the selective least squares estimates in (4) are not consistent.

Let $X_i = (1, X_{i1}, \ldots, X_{ip})'$, $\alpha = (a_0, a_1, \ldots, a_p)'$ and $\beta = (b_0, b_1, \ldots, b_p)'$. Suppose we have initial coefficient estimates α_0 and β_0 such that $|\alpha_0 - \alpha| < v$ and $|\beta_0 - \beta| < v$ where v is a $(p + 1)$-dimensional positive vector. We assume the $(i + 1)$st element of v is less than $|a_i - b_i|/4$ for $i = 0, 1, \ldots, p$.

For $M > 0$, let

$$\mathcal{A}(M) = \{i \text{ s.t. } |X_i'(\alpha_0 - \beta_0)| > M + 2|X_i|'v\}$$

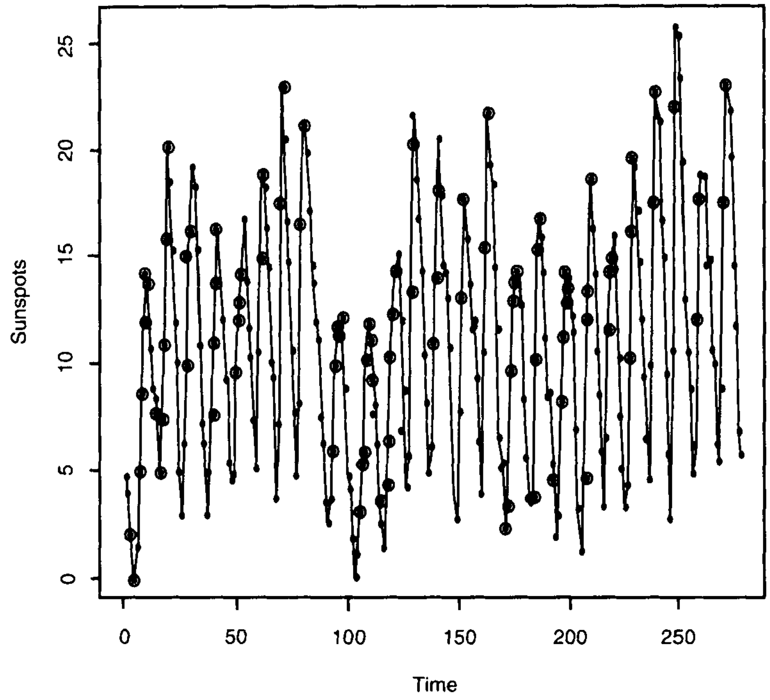


FIGURE 9. Time plot of the sunspot series. Circled points belong to the first regime in the open-loop threshold model.

where $|X_i| = (1, |X_{i1}|, \ldots, |X_{ip}|)'$. In addition, let

$$T_1(M) = \{i, \text{ s.t. } (Y_i - X_i'\alpha_0)^2 < (Y_i - X_i'\beta_0)^2, i \in \mathcal{A}(M)\}$$

$$T_2(M) = \{i, \text{ s.t. } (Y_i - X_i'\alpha_0)^2 > (Y_i - X_i'\beta_0)^2, i \in \mathcal{A}(M)\}.$$

Let $\hat{\alpha}(M)$ be the ordinary least squares estimator of α using the observations in set $T_1(M)$ and $\hat{\beta}(M)$ be that of β using set $T_2(M)$. We have the following theorem.

THEOREM 1. *Assume (Y_i, X_i, I_i) follow a digression model in (2). The sequence (X_i, I_i) is a finite-order Markov chain and is stationary and ergodic. The marginal density functions of the X_{ji}s are strictly positive on the real line. Assume $\Sigma_k(M) = E\{XX'|I = k, \mathcal{A}(M)\}$, $k = 1, 2$, are finite for any M and $\|\Sigma_1(M)^{-1}\Sigma_2(M)\|$ and $\|\Sigma_2(M)^{-1}\Sigma_1(M)\|$ are bounded uniformly for all M. Assume $0 < h_1 < P\{I_i = 1|i \in \mathcal{A}(M)\} < h_2 < 1$ for all M and some h_1 and h_2. Then for any $\varepsilon > 0$, there is an $M > 0$ and an $N > 0$ such that if the sample size is greater than N, the discarding least squares estimators $\hat{\alpha}(M)$ and $\hat{\beta}(M)$ with the constant M satisfy*

$$\|\hat{\alpha}(M) - \alpha\| < \varepsilon \qquad \|\hat{\beta}(M) - \beta\| < \varepsilon \qquad \text{a.s.}$$

PROOF. For $i \in \mathcal{A}(M)$,

$$\begin{aligned} P\{i \in T_1(M)|I_i = 2, i \in \mathcal{A}(M)\} \\ &= P\{(Y_i - X_i'\alpha_0)^2 < (Y_i - X_i'\beta_0)^2|I_i = 2, i \in \mathcal{A}(M)\} \\ &= P[X_i'(\beta_0 - \alpha_0)\{2\varepsilon_i^{(2)} + X_i'(2\beta - \alpha_0 - \beta_0)\} < 0|i \in \mathcal{A}(M)] \end{aligned}$$

$$= P\{2\varepsilon_i^{(2)} < -|X_i'(2\beta - \alpha_0 - \beta_0)|\,|i \in \mathcal{H}(M)\} \tag{10}$$

$$\leqslant P(\varepsilon_i^{(2)} < -M/2) \tag{11}$$

where (10) is due to the fact that $X_i'(2\beta - \alpha_0 - \beta_0)$ has the same sign as $X_i'(\beta_0 - \alpha_0)$ since $X_i'(2\beta - \alpha_0 - \beta_0) = X_i'(\beta_0 - \alpha_0) + 2X_i'(\beta - \alpha_0)$ and $|X_i'(\beta_0 - \alpha_0)| \geqslant |2X_i'(\beta - \alpha_0)|$ for $i \in \mathcal{H}(M)$. Equation (11) is due to

$$\begin{aligned} |X_i'(2\beta - \alpha_0 - \beta_0)| &\geqslant |X_i'(\beta_0 - \alpha_0)| - 2|X_i'(\beta - \beta_0)| \\ &> M + 2|X_i'|\nu - 2|X_i'(\beta - \beta_0)| > M. \end{aligned}$$

Let $S_1(M) = \{i, \text{ s.t. } i \in \mathcal{H}(M),\ I_i = 1\}$ and $S_2(M) = \{i, \text{ s.t. } i \in \mathcal{H}(M),\ I_i = 2\}$. For classification $T_1(M)$, let X_{T_1} be the explanatory matrix using the observations in $T_1(M)$. Let I_{S_k}, $k = 1, 2$, be a diagonal matrix with ith diagonal element being 1 if the ith observation in $T_1(M)$ is in S_k and 0 otherwise (note $I_{S_1} + I_{S_2}$ is the identity matrix), then the least squares estimate of α using $T_1(M)$ is

$$\begin{aligned} \hat{\alpha} &= (X_{T_1}'X_{T_1})^{-1}X_{T_1}'Y_{T_1} \\ &= (X_{T_1}'X_{T_1})^{-1}X_{T_1}'(I_{S_1}X_{T_1}\alpha + I_{S_2}X_{T_1}\beta + I_{S_1}\varepsilon^{(1)} + I_{S_2}\varepsilon^{(2)}) \\ &= (X_{T_1}'X_{T_1})^{-1}X_{T_1}'\{X_{T_1}\alpha + I_{S_2}X_{T_1}(\beta - \alpha) + I_{S_1}\varepsilon^{(1)} + I_{S_2}\varepsilon^{(2)}\} \\ &= \alpha + (X_{T_1}'X_{T_1})^{-1}(X_{T_1}'I_{S_2}X_{T_1})(\beta - \alpha) + (X_{T_1}'X_{T_1})^{-1}X_{T_1}'\{\varepsilon^{(1)} + I_{S_2}(\varepsilon^{(2)} - \varepsilon^{(1)})\} \\ &= \alpha + E_1 + E_2. \end{aligned}$$

Let $n_k = \sum_i \delta(i \in S_k)$ be the number of observations in S_k for $k = 1, 2$ and $\Sigma_k(M) = E\{X_i'X_i|I_i = k,\ i \in \mathcal{H}(M)\}$, $k = 1, 2$. By the assumptions, as the sample size n goes to infinity, n_k, $k = 1, 2$, goes to infinity. Denote $\delta(\cdot)$ as the indicator function. Then, by the law of large numbers, under the ergodic condition,

$$\begin{aligned} \frac{(X_{T_1}'I_{S_1}X_{T_1})/n}{n_1/n} &\rightarrow E\{X_i'X_i\delta(i \in T_1)|I_i = 1\} \\ &= E\{X_i'X_i - X_i'X_i\delta(i \in T_2)|I_i = 1\} \\ &= \Sigma_1(M) - E[X_i'X_i\delta\{2\varepsilon_i^{(1)} < -|X_i'(2\alpha - \alpha_0 - \beta_0)|\}|I_i = 1] \\ &= \Sigma_1(M) - \Delta_1 \end{aligned}$$

where

$$\begin{aligned} \|\Delta_1\| &= \|E[X_i'X_i\delta\{2\varepsilon_i^{(1)} < -|X_i'(2\alpha - \alpha_0 - \beta_0)|\}|I_i = 1]\| \\ &\leqslant \|E\{X_i'X_i\delta(\varepsilon_i^{(1)} < -M/2)|I_i = 1\}\| \\ &= \|E(X_i'X_i|I_i = 1)\|P(\varepsilon_i^{(1)} < -M/2) \\ &= \|\Sigma_1(M)\|P(\varepsilon_i^{(1)} < -M/2). \end{aligned}$$

Similarly,

$$\begin{aligned} \|\Delta_2\| = \|(X_{T_1}'I_{S_2}X_{T_1})/n_2\| &\rightarrow \|E\{X_i'X_i\delta(i \in T_1)|I_i = 2\}\| \\ &= \|E\{X_i'X_i\delta\{2\varepsilon_i^{(2)} < -|X_i'(2\beta - \alpha_0 - \beta_0)|\}|I_i = 2]\| \\ &\leqslant \|E\{X_i'X_i\delta(\varepsilon_i^{(2)} < -M/2)|I_i = 2\}\| \\ &= \|E\{X_i'X_i|I_i = 2)\|P(\varepsilon_i^{(2)} < -M/2) \\ &= \|\Sigma_2(M)\|P(\varepsilon_i^{(2)} < -\ M/2). \end{aligned}$$

Hence,

$$\frac{X'_{T_1}X_{T_1}}{n_1} = \frac{X'_{T_1}I_{S_1}X_{T_1}}{n_1} + \frac{(n_2/n_1)X'_{T_1}I_{S_2}X_{T_1}}{n_2} \to \Sigma_1(M) - \Delta_1 + \Delta_2\left(\frac{n_2}{n_1}\right)$$

where $\|\Delta_1\| \leqslant \|\Sigma_1(M)\| P(\varepsilon_i^{(1)} < -M/2)$ and $\|\Delta_2\| \leqslant \|\Sigma_2(M)\| P(\varepsilon_i^{(2)} < -M/2)$. By the assumptions, we have

$$\|E_1\|^2 \leqslant \left(\frac{n_2}{n_1}\right)^2 \left\| \frac{(X'_{T_1}X_{T_1}/n_1)^{-1}(X'_{T_1}I_{S_2}X_{T_1})}{n_2} \right\|^2 \|\alpha - \beta\|^2 < \varepsilon/2$$

for large enough sample size n and the constant M.

On the other hand, for fixed M,

$$\begin{aligned}\|E_2\|^2 &= \|(X'_{T_1}X_{T_1})^{-1}\|\sigma_1^2 + \|(X'_{T_1}X_{T_1})^{-1}X'_{T_1}I_{S_2}X_{T_1}(X'_{T_1}X_{T_1})^{-1}\|(\sigma_1^2 + \sigma_2^2) \\ &= \frac{\sigma_1^2}{n_1}\left\|\left(\frac{X'_{T_1}X_{T_1}}{n_1}\right)^{-1}\right\| + \frac{n_2}{n_1^2}(\sigma_1^2 + \sigma_2^2)\left\|\left(\frac{X'_{T_1}X_{T_1}}{n_1}\right)^{-1}\left(\frac{X'_{T_1}I_{S_2}X_{T_1}}{n_2}\right)\left(\frac{X_{T_1}X_{T_1}}{n_1}\right)^{-1}\right\| \\ &< \varepsilon/2\end{aligned}$$

for large enough n.

Hence, for any $\varepsilon > 0$, we can find large enough M and large enough n such that $\|E_1\| < \varepsilon/2$ and $\|E_2\| < \varepsilon/2$. The theorem follows. □

Next we argue that the selective least squares estimators in (4) are not consistent.

THEOREM 2. *Suppose $\hat{\alpha}$ and $\hat{\beta}$ are the least squares estimators and W_1 and W_2 are the partition of $\{1, \ldots, n\}$ such that the global minimum of* (4) *is achieved. Under the assumption of Theorem* 1, *$\hat{\alpha}$ and $\hat{\beta}$ are not consistent estimators.*

PROOF. Let α and β be the true coefficients. Note that if $I_i = 1$,

$$\begin{aligned}P\{(Y_i - X_i'\alpha)^2 > (Y_i - X_i'\beta)^2\} &= P[(\varepsilon_i^{(1)})^2 > \{\varepsilon_i^{(1)} + X_i'(\alpha - \beta)\}^2] \\ &= P[X_i'(\alpha - \beta)\{2\varepsilon_i^{(1)} + X_i'(\alpha - \beta)\} > 0]. \qquad (12)\end{aligned}$$

It is easy to see that for a small $\delta_1 > 0$, there is a set Ω in $\mathcal{R}^p$ with positive measure with respect to X'such that (12) is greater than δ_1 uniformly. Since (12) is continuous in $\alpha - \beta$, if the least square estimators $\hat{\alpha}$, $\hat{\beta}$ are consistent, we can find an N such that for the sample size n greater than N,

$$P(i \in W_2 | I_i = 1) = P\{(Y_i - X_i'\hat{\alpha})^2 > (Y_i - X_i'\hat{\beta})^2 | i \in S_1\} > \delta_1/2$$

uniformly for $X_i \in \Omega$. Then by a similar argument in the proof of Theorem 1, as n goes to ∞, we have

$$\|\hat{\alpha} - \alpha\| > \delta\|\alpha - \beta\|$$

which contradicts the consistency of $\hat{\alpha}$. □

ACKNOWLEDGEMENTS

This research was supported in part by NSF grant DMS-9301193 and Texas A&M University Department of Statistics. The author thanks an anonymous referee for helpful comments and suggestions.

REFERENCES

Chen, R. and Liu, J. S. (1995) Predictive updating methods with application to Bayesian classification. *J. R. Statist. Soc. B.* in press.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* 85, 398–409.

Ghaddar, D. K. and Tong, H. (1981) Data transformation and self-exciting threshold autoregression. *Appl. Stat.* 30, 238–48.

Goldfeld, S. M. and Quandt, R. E. (1973) A Markow model for switching regression. *J. Econometrics* 1, 3–16.

Haggan, V. and Ozaki, T. (1981) Modeling nonlinear vibration using an amplitude-dependent autoregressive time series model. *Biometrika* 68, 189–96.

Izenman, A. J. (1983) J. R. Wolf and J. A. Wolfer: An historical note on the Zurich sunspot relative numbers. *J. R. Statist. Soc.* A, 146, 311–18.

Kotz, S., Johnson, N. and Read, C. (1988) *Encyclopedia of Statistical Sciences*. New York: Wiley.

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

Liu, J. S., Wong, W. H. and Kong, A. (1995) Correlation structure and convergence rate of the Gibbs sampler with various scans. *J. R. Statist. Soc. B*, 37, 157–170.

Priestley, M. B. (1980) State-dependent models: a general approach to non-linear time series analysis. *J. Time Series Anal.* 1, 47–71.

— (1988) *Non-linear and Non-stationary Time Series Analysis*. New York: Academic Press.

Quandt, R. E. and Ramsey, J. B. (1978) Estimating mixtures of normal distributions and switching regressions. *J. Am. Statist. Assoc.* 73, 730–38.

Shumway, R. H. and Stoffer, D. S. (1991) Dynamic linear models with switching. *J. Am. Statist. Assoc.* 86, 763–69.

Subba Rao, T. and Gabr, M. M. (1984) An introduction to bispectral analysis and bilinear time series models. New York: Springer-Verlag.

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

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

Tsay, R. S. (1989) Testing and modeling threshold autoregressive processes. *J. Am. Statist. Assoc.* 84, 231–40.