**Nonlinear Additive ARX Models**

Rong Chen; Ruey S. Tsay

*Journal of the American Statistical Association*, Vol. 88, No. 423 (Sep., 1993), 955-967.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199309%2988%3A423%3C955%3ANAAM%3E2.0.CO%3B2-3

*Journal of the American Statistical Association* is currently published by American Statistical Association.

# Nonlinear Additive ARX Models

RONG CHEN and RUEY S. TSAY*

We consider in this article a class of nonlinear additive autoregressive models with exogenous variables for nonlinear time series analysis and propose two modeling procedures for building such models. The procedures proposed use two backfitting techniques (the ACE and BRUTO algorithms) to identify the nonlinear functions involved and use the methods of best subset regression and variable selection in regression analysis to determine the final model. Simulated and real examples are used to illustrate the analysis.

KEY WORDS: Additivity; Alternating conditional expectation (ACE) algorithm; Best subset regression; BRUTO algorithm; River flow; Time series; Variable selection.

## 1. INTRODUCTION

Nonlinear time series analysis has gained much attention in recent years, due primarily to the fact that linear time series models have encountered various limitations in real applications and modern computers have provided advanced computational power that makes possible the nonlinear analysis. In addition, the development in nonparametric regression has established a solid foundation for nonlinear time series analysis.

Many nonlinear time series models have been introduced in the literature and shown to be useful in some applications. For instance, Tong (1978, 1990) proposed the threshold autoregressive model and demonstrated that the model is capable of describing the asymmetric limit cycle of the annual sunspot number, and Haggan and Ozaki (1981) considered the exponential autoregressive model and showed that the model is useful in modeling sound vibration. But most nonlinear models use explicit parametric forms that can, at best, be regarded as rough approximations to the underlying nonlinear characteristics of interest. It is usually hard to justify a priori the appropriateness of such an explicit model in real applications. To overcome this justification problem and to make use of recent developments in nonparametric regression, researchers in nonlinear time series analysis began to explore the possibility of using data-driven methods, such as nonparametric density estimation, to identify the underlying characteristics of a time series. For example, Robinson (1983) investigated asymptotic properties of nonparametric density estimation for time series data; Auestad and Tjøstheim (1990) applied a multivariate kernel smoothing method to estimate the conditional mean and conditional variance of a nonlinear autoregression; Lewis and Stevens (1991) used the multivariate adaptive regression splines (MARS) of Friedman (1991) to build adaptive spline threshold autoregressive models; and Chen and Tsay (1993) used an arranged local regression procedure to construct functional-coefficient autoregressive models.

In this article we also adopt certain ideas of nonparametric regression for nonlinear time series analysis. Our objective is to study a class of nonlinear additive autoregressive (NAAR) model with exogenous variables and to consider procedures for building such models. In particular, we use two backfitting algorithms in nonparametric regression to estimate the unknown functions. These backfitting algorithms are the alternating conditional expectation (ACE) algorithm of Breiman and Friedman (1985) and the BRUTO algorithm of Hastie and Tibshirani (1990). Both algorithms use ideas of cross-validation in selecting the smoothing parameter and appear to work reasonably well for stationary time series. For model building, we use the idea of best subset regression in multiple linear regression analysis to select the lagged variables of an NAAR model. Experience shows that such a modeling procedure is promising.

Because both ACE and BRUTO algorithms are nonparametric tools, our analysis is nonparametric in nature. But in some cases we treat the nonparametric procedure as a preliminary data analysis on which a parametric model can be constructed. Parametric models have some advantages in applications. First, they are easier to understand and interpret. Second, they can simplify forecasts (e.g., obtaining forecast intervals). Third, model comparison in parametric context has been well studied, so the difficulty of model comparison encountered in using nonparametric tools can be avoided.

This article is organized as follows. Section 2 introduces the nonlinear additive models used in the article. Section 3 reviews the basic idea of backfitting procedures in nonparametric regression, especially the ACE algorithm and the BRUTO algorithm. Section 4 proposes two modeling procedures. The best subset modeling procedure is analogous to the best subset regression in linear regression analysis. The only difference is that the former uses the ACE algorithm in estimation. The other modeling procedure proposed uses the BRUTO algorithm, which is an adaptive backfitting procedure, and uses generalized cross-validation to select significant explanatory variables. This latter procedure thus automatically selects a final model within a given class of candidate models. Section 5 illustrates the two modeling procedures by analyzing both real and simulated examples.

## 2. NONLINEAR ADDITIVE MODELS

The NAAR model considered in this article can be written as

$$y_t = f_1(y_{t-i_1}) + f_2(y_{t-i_2}) + \cdots + f_p(y_{t-i_p}) + \varepsilon_t, \quad (1)$$

* Rong Chen is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. Ruey S. Tsay is Professor of Statistics, Graduate School of Business, University of Chicago, IL 60637. This research was supported in part by National Science Foundation Grant DMS-9103250, the University of Chicago Graduate School of Business, and the Texas A&M University Department of Statistics. The authors thank H. Tong for the riverflow data and an associate editor and two anonymous referees for their helpful comments.


Journal of the American Statistical Association
September 1993, Vol. 88, No. 423, Theory and Methods

where $\{\varepsilon_t\}$ is a sequence of iid random variables, $i_j$'s are positive integers, and $f_i(\cdot)$'s are measurable real-valued functions. In applications, we further assume that $f(\cdot)$'s are smooth functions. This model is a simple generalization of the first-order nonlinear autoregressive model of Jones (1978), and is a time series counterpart of the generalized additive model of Hastie and Tibshirani (1991) in regression analysis. The key feature of the model is additivity (i.e., no interactions between different lagged variables), so that each function is univariate. This feature substantially simplifies the complexity involved in empirical model building.

On the other hand, one may argue that the additivity assumption of model (1) is strong and might hinder the model's applicability. We justify using the model on several grounds. First, it provides a sensible alternative to the general nonlinear autoregressive model

$$y_t = f(y_{t-1}, \ldots, y_{t-p}) + \varepsilon_t, \tag{2}$$

which often encounters the difficulty known as the "curse of dimensionality" in applications. Because the nonlinear function in (2) is multidimensional, analysis of such a model often requires multivariate smoothing. The virtue of nonparametric smoothing is to make use of "local properties" of the data; for a multivariate problem, a large sample is needed to obtain reliable local estimates. Consequently, for the sample sizes commonly encountered in practice, nonparametric estimates of model (2) often associate with large variations, especially when $p$ is not small. See figure 4.1 of Hastie and Tibshirani (1990, p. 84) for a simple yet informative illustration of the curse of dimensionality. The NAAR model in (1) uses only univariate smoothing, which is much easier and better understood. Second, the NAAR model is sufficiently flexible. It encompasses linear autoregressive models and many interesting nonlinear models as special cases. For instance, linear combinations of simple trigonometric series are NAAR models. Finally, the validity of the additivity assumption can be checked in applications. If necessary, interaction terms can be incorporated approximately into the model by treating them as pseudoexogenous variables.

When exogenous variables are available, we extend further the model to a class of NAAR$X$ models, where $X$ stands for exogenous variables. An NAAR$X$ model with an exogenous variable can be written as

$$y_t = f_1(y_{t-i_1}) + \cdots + f_p(y_{t-i_p}) + g_1(x_{t-j_1}) + \cdots + g_q(x_{t-j_q}) + \varepsilon_t, \tag{3}$$

where $g_i(\cdot)$'s are measurable real-valued functions, $j_k$'s are nonnegative integers, and the exogenous series $\{x_t\}$ is independent of the noise series $\{\varepsilon_t\}$. Again, the additivity of the lagged variables of $x_t$ is used to reduce the dimensionality of the model and can be checked in applications.

## 3. NONPARAMETRIC ESTIMATION

In this section we briefly review the technique of nonparametric smoothing and two backfitting algorithms used in this article. These technique and algorithms are the basic tools we used to estimate the unknown functions $f(\cdot)$'s and $g(\cdot)$'s of an NAAR$X$ model.

### 3.1 Scatterplot Smoothers

Consider a generic regression model $Y_i = f(X_i) + \varepsilon_i$, where $f(\cdot)$ is a unknown smooth function. The basic idea of nonparametric estimation of $f(\cdot)$ is to apply a certain smoother to the scatterplot of the bivariate data $(x_i, y_i)$ for $i = 1, \ldots, n$. The scatterplot smoother used in the ACE algorithm is the supersmoother of Friedman and Stuetzle (1982). Briefly speaking, the building block of supersmoother is a symmetric $k$ nearest-neighbor linear least squares procedure. That is, for each $x_j$, the nearest $k/2$ of the $x_i$ to the right and the nearest $k/2$ of the $x_i$ to the left of $x_j$ are used along with $x_j$ and the corresponding $y_i$'s to fit a straight line by least squares. This line is used as the fit for the region from $x_j$ to half the distance to each of its nearest neighbors. The value of $k$ is chosen for each $x_j$ using a local cross-validation technique, a leave-one-out cross-validation that used only neighboring points of $x_j$. Details of the local cross-validation can be found in the above reference.

For the BRUTO algorithm, various smoothers can be used. In this article we use mainly a Gaussian kernel smoother. More specifically, an estimate $\hat{f}(x)$ of $f(x)$ is computed by

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^n K[(x - x_i)/\lambda] y_i}{\sum_{i=1}^n K[(x - x_i)/\lambda]},$$

where $K[\cdot]$ is a Gaussian kernel and the smoothing parameter $\lambda$ is chosen by minimizing a modified generalized cross-validation (GCV) criterion. Focusing on the observed values $x_1, \ldots, x_n$, we can write the fitted vector of the smoother as $\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$, where $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{S}_\lambda$ is the $n \times n$ smoother matrix associated with the smoothing parameter $\lambda$. The GCV criterion is then to select $\lambda$ that minimizes the function

$$\mathrm{GCV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - \mathrm{tr}(\mathbf{S}_\lambda)/n} \right\}^2, \tag{4}$$

where $\mathrm{tr}(\mathbf{S}_\lambda)$ denotes the trace of $\mathbf{S}_\lambda$. In the BRUTO algorithm, some modification is made to simplify the computation of $\mathrm{tr}(\mathbf{S}_\lambda)$; see Section 3.4.

### 3.2 A Backfitting Procedure

The NAAR$X$ model often involves several explanatory variables, each with an unknown functional form. Therefore, it is necessary to implement the previously described methods of nonparametric estimation in a systematic manner. To this end, we follow Breiman and Friedman (1985) and Hastie and Tibshirani (1990, p. 91) by using a backfitting procedure in which the individual functions are estimated iteratively, conditioned on the results of the other functions. Consider the nonlinear regression model

$$Y = \sum_{i=1}^p f_i(X_i) + \varepsilon, \tag{5}$$

where $f_i(\cdot)$'s are measurable smooth functions and $\varepsilon$ is independent noise. The main idea of backfitting is that if the additive model is correct, then for any $i$ we have

$$f_i(X_i) = E\left[Y - \sum_{j \neq i} f_j(X_j) \mid X_i\right].$$

Consequently, to estimate $f_i(\cdot)$ we can treat $Y - \sum_{j \neq i} f(X_j)$ as the conditional response variable and use the scatterplot smoothers mentioned earlier. In practice, all $f_i(\cdot)$'s are unknown, so that the estimates are iterated until they all converge. In summary, the backfitting procedure used can be described as follows:

1. Initialize: $f_j(X_j) = f_j^0(X_j), j = 1, \ldots, p$.
2. Cycle: $j = 1, \ldots, p, 1, \ldots, p, \ldots$

$$f_j(X_j) = S_j\left(Y - \sum_{k \neq j} f_k\left(X_k\right) \mid X_j\right),$$

where $S_j(W \mid Z)$ denotes a smoothing estimate of the variable $W$ using the explanatory variable $Z$.

3. Continue Step 2 until all of the individual functions converge.

### 3.3 The ACE Algorithm

The ACE algorithm is well known and can be found in Breiman and Friedman (1985) and Hastie and Tibshirani (1990, p. 178). The algorithm has been shown to converge under certain conditions. For instance, if $f_i(\cdot)$'s are jointly normal and $\varepsilon_t$ is also normal and is independent of the explanatory variables, then the ACE estimates converge as the number of iterations increases. (See also Koyak 1990 for related work.) In our implementation we do not transform the dependent variable $y_t$, and we have adjusted the algorithm accordingly.

### 3.4 The BRUTO Algorithm

The BRUTO algorithm combines backfitting and adaptive smoothing parameter selection. It also allows for a null fit, which enables us to remove the associated explanatory variable from further consideration. This feature is particularly useful in lag selection of NAAR$X$ models. For the general nonlinear regression model in (5), the modified GCV criterion used in the BRUTO algorithm is

$$\text{GCV}^b(\lambda_1, \ldots, \lambda_p) = \frac{\sum_{i=1}^n [y_i - \sum_j \hat{f}_{j,\lambda_j}(x_{ij})]^2}{n\{1 - [1 + \sum_{j=1}^p (\text{tr } \mathbf{S}_j(\lambda_j) - 1)]/n\}^2},$$

where $n$ is the sample size, $x_{ij}$ denotes the $i$th observation of the variable $X_j$, and $\mathbf{S}_j(\lambda_j)$ is the smoother matrix with smoothing parameter $\lambda_j$ for the $j$th explanatory variable. This is slightly different from the usual GCV when $p \geq 2$. The modification is based on computational consideration and involves only the function $1 + \sum_{j=1}^p [\text{tr } \mathbf{S}_j(\lambda_j) - 1]$ in the denominator, which is an approximation of what is commonly referred to as the degrees of freedom of the smoothers used and can be interpreted as the penalties for using the explanatory variables. With normal density kernel, the trace of the smoother matrix becomes

$$\text{tr } \mathbf{S}_j(\lambda_j) = \sum_{i=1}^n \frac{K(0)}{\sum_{v=1}^n K[(x_{vj} - x_{ij})/\lambda_j]},$$

where $K(\cdot)$ is the kernel function. This can be easily evaluated. We also include the null fit (indicated by smoothing parameter $\lambda_j = 0$) and the linear fit (indicated by $\lambda_j = -1$) in the procedure.

In the BRUTO algorithm an iteration contains the selection of the parameters $\lambda_j$ for $j = 1, \ldots, p$. This is carried out one parameter at a time by applying the appropriate smoother to the transformed response, say $y_t - \sum_{k \neq j} \hat{f}_{k,\lambda_k}(x_{ik})$ for $\lambda_j$, but selecting the smoothing parameter to minimize the global criterion $\text{GCV}^b(\lambda_1, \ldots, \lambda_p)$. More specifically, having obtained the best candidate for each of the $p$ parameters, we incorporate only the update that corresponds to the minimum $\text{GCV}^b(\lambda_1, \ldots, \lambda_p)$. Therefore, each iteration updates only one transformation. The iteration is continued until $\text{GCV}^b(\lambda_1, \ldots, \lambda_p)$ converges. The convergence must occur, because each iteration produces a decrease in the criterion.

The kernel smoother used in the BRUTO algorithm can be replaced by locally weighted lines. This approach will automatically include the linear fit and may ease the bias of kernel estimates at the end points. But our limited experience shows that the results of locally weighted lines do not differ much from that of the kernel estimates. Further, it is slightly more difficult to compute the approximate degrees of freedom in using locally weighted lines.

### 3.5 Remarks

There are some potential problems in applying the two backfitting procedures to time series data. Unlike regression analysis, time series data are serially correlated. In applications, strong serial dependence might mislead the procedures to produce erroneous transformations. For instance, if $y_t$ is close to unit-root nonstationarity in the sense that its lag-1 serial correlation is close to unity, then the ACE algorithm tends to suggest linear transformation for $y_{t-1}$. Such misleading results will not occur asymptotically, provided that the serial dependence decays sufficiently fast. Robinson (1983) and Auestad and Tjøstheim (1990) showed that under the $\alpha$-mixing condition that holds generally under geometrical ergodicity, nonparametric regression techniques such as kernel smoothing are applicable to time series data. For NAAR models, some sufficient conditions of geometrical ergodicity are available in Chen (1990). In particular, if we treat NAAR models as special cases of functional-coefficient autoregressive models, then results of geometrical ergodicity in Chen and Tsay (1993) apply.

Hart and Vieu (1990) showed that the leave-one-out cross-validation to bandwidth selection in kernel density estimation continues to be asymptotically optimal under a strong mixing condition. These authors also found via simulations that improving on bandwidth selection by ordinary cross-validation is possible in finite-size samples when the serial correlation is strong.

The literature on nonparametric regression for dependent data is relatively sparse. The results just cited indicate that the ACE and BRUTO algorithms are still applicable, provided that the serial dependence is not strong. In this article we use these algorithms as exploratory tools for model building. The models built need to be checked. In this sense the efficiency of the algorithms is not too critical. Of course, further study on the use of ACE and BRUTO algorithms in time series analysis is needed to justify the general use of the proposed modeling procedures.

## 4. MODELING PROCEDURES

Here we consider two modeling procedures for NAAR$X$ models. The NAAR models can be handled easily by dropping the exogenous variable. The first modeling procedure uses the idea of best subset regression and the technique of ACE; the second procedure uses the BRUTO algorithm. To illustrate the performance of the proposed procedures, we conduct a simulation study on the key step of each procedure.

### 4.1 The Best Subset Modeling Procedure

This procedure is analogous to the best subset regression procedure in linear regression analysis and consists of three steps. The first step is to determine the maximum order of $i_p$ and $j_q$ in Model (3). This is done by computing the estimated residual variance of a NAAR$X(l, l)$ model $\hat{\sigma}_l^2$ for $l = 1, 2, \ldots$. Here the order $(l, l)$ denotes a NAAR$X$ model in (3) with $i_j = j$ for $j = 1, \ldots, l$ and $j_k = k$ for $k = 0, \ldots, l$, and the estimation is done by the ACE algorithm. In other words, we use the ACE algorithm to fit a full NAAR$(l, l)$ model and obtain the associated residual variance $\hat{\sigma}_l^2$ for $l = 1, \ldots, L$, where $L$ is a prespecified positive integer. We then plot $\hat{\sigma}_l^2$ against $l$ and select the order $p$ such that the residual variances show no major reduction after $\hat{\sigma}_p^2$.

The second step is to identify the best subset models. For convenience, we standardize the $y_t$ series in this step. That is, let $y_t^*$ be the standardized series of $y_t$ such that $E(y_t^*) = 0$ and $\text{var}(y_t^*) = 1$. For a given size $i + j$ such that $1 \le i + j \le 2p$, where $p$ is the maximum order selected in the first step, we consider all possible nonlinear autoregressions with exogenous variables of size $i + j$, say

$$y_t^* = f_1(y_{t-u_1}) + \cdots + f_i(y_{t-u_i}) + g_1(x_{t-v_1}) + \cdots + g_j(x_{t-v_j}) + \varepsilon_t$$

where $\{u_1, \ldots, u_i\}$ and $\{v_1, \ldots, v_j\}$ are subsets of $\{1, \ldots, p\}$ and $\{0, \ldots, p\}$. Again, the estimation is done by the ACE algorithm. We then select the best nonlinear model of size $i + j$ by choosing the one that maximizes

$$R^2 = 1 - \sum_{t=1}^{n} \left[ y_t^* - \sum_{k=1}^{i} \hat{f}_k(y_{t-u_k}) - \sum_{k=1}^{j} \hat{g}_k(x_{t-v_k}) \right]^2,$$

where $\hat{f}_k(\cdot)$ and $\hat{g}_k(\cdot)$ are estimates of $f_k(\cdot)$ and $g_k(\cdot)$. This $R^2$ is in fact the criterion used in the ACE algorithm.

Finally, in the third step we compare the best subset regressions for different sizes $i + j$ and select the model that best fits the data. This final step can be achieved in various ways. For instance, one can plot the $R^2$ against the size $i + j$ and select a model based on the shape of the plot. Another possibility is to build a parametric model for each size $i + j$ based on $\hat{f}_k(y_{t-u_k})$ and $\hat{g}_k(x_{t-v_k})$ of the best subset regression and then use some well-known information criterion such as Akaike's information criterion (AIC) (Akaike 1974) to select the best parametric model. In this article we plot the $R^2$ and select a final model based on the shape of the plot.

Obviously, the key step of the proposed best subset modeling procedure is the second step. To illustrate the performance of this step, we conduct a simulation study. Consider the following six models:

1. $y_t = .8y_{t-1} + \varepsilon_t$,
2. $y_t = .8y_{t-1} - .6y_{t-2} + \varepsilon_t$,
3. $y_t = .8y_{t-1} - .6y_{t-3} + \varepsilon_t$,
4. $y_t = .8 \log(1 + 3y_{t-1}^2) - .6 \log(1 + 3y_{t-3}^2) + \varepsilon_t$,
5. $y_t = 1.5 \sin((\pi/2)y_{t-2}) - 1.0 \sin((\pi/2)y_{t-3}) + \varepsilon_t$,
6. $y_t = (.5 - 1.1 \exp(-50y_{t-1}^2))y_{t-1} + (.3 - .5 \exp(-50 \times y_{t-3}^2))y_{t-3} + \varepsilon_t$,

where $\{\varepsilon_t\}$ is a sequence of iid $N(0, 1)$ random variates. For each model we generated 20 realizations of size 300 and applied the best subset modeling procedure to the data. The maximum order of $i_p$ is fixed at 5. We then enumerated the number of realizations for which the "true" model is selected as the best subset of the appropriate size. It turns out that for the first five models the "true" model is always selected as the best subset model of the appropriate size. For Model 6, the successful rate is 16 out of 20. This encouraging result suggests that the best subset modeling procedure could be useful in modeling NAAR and NAAR$X$ models.

### 4.2 An Alternative Modeling Procedure

Now we consider another procedure for modeling the NAAR$X$ models. This alternative procedure uses the adaptive backfitting BRUTO algorithm of Hastie (1989) for additive regressions to determine the order and lags of an NAAR$X$ model. The proposed approach assumes that the maximum orders of $i_v$ and $j_k$ in (3) are given. These maxima can be prespecified positive integers or can be determined by the method discussed in the first step of Section 4.1. With the maximum orders given, the BRUTO algorithm is then used directly to select the lag indices $i_v$ and $j_k$ and to obtain estimates of the associated functions $f_{i_v}(\cdot)$ and $g_{j_k}(\cdot)$. Because the algorithm allows for null fit, it behaves like the stepwise procedure for variable selection in multiple linear regression analysis.

Define the weight of an explanatory variable, $y_{t-i}$ or $x_{t-j}$, as the increment in GCV [b] when that variable is removed from the model. This weight plays an important role in using the result of the BRUTO algorithm. Experience shows that the weights of some explanatory variables can be small, but not 0. In this situation further investigation is needed to avoid misinterpreting the result of the algorithm.

To experience the BRUTO algorithm, we simulated 15 series from each of the following NAAR models with sample

size 200:

1. $y_t = [.8 - 1.1 \exp(-50y_{t-1}^2)]y_{t-1} + \varepsilon_t$,
2. $y_t = [.5 - 1.1 \exp(-50y_{t-1}^2)]y_{t-1} + [.3 - .5 \exp(-50 \times y_{t-3}^2)]y_{t-3} + \varepsilon_t$,
3. $y_t = -.3y_{t-1}I(y_{t-1} < 0) + .8y_{t-1}I(y_{t-1} \geq 0) + \varepsilon_t$,
4. $y_t = .8y_{t-1} + \varepsilon_t$,

where, again, $\{\varepsilon_t\}$ is a sequence of iid $N(0, 1)$ random variates. We then applied the BRUTO algorithm to select a final model. With the maximum order fixed at 5, the selection results of the algorithm are

| | lags | | | | | | |
|---|---|---|---|---|---|---|---|
| model | 1 | 1, 2 | 1, 3 | 1, 4 | 1, 5 | 1, 3, 4 | 1, 3, 5 |
| (1) | 9* | 4 | 2 | | | | |
| (2) | | | 14* | | | | 1 |
| (3) | 8* | 1 | 1 | 2 | 2 | 1 | |
| (4) | 10* | | 2 | 2 | 1 | | |

where *'s indicate the "true model" of the series. In this study, we removed from the final selection the explanatory variables whose weight is less than 1% of the final GCV. The algorithm appears to work well.

### 4.3 Some Discussions

The best subset modeling procedure requires some subjective judgment in order determination, especially in the first step. But selecting the maximum order in this step is not critical. We do so mainly to reduce the number of possible subsets in the second step and hence to simplify the computation involved in model search. One can take a conservative approach by selecting a relatively large maximum order, provided that computation is not a serious limitation.

In linear regression analysis, efficient algorithms are available to evaluate the best subset regressions; see, for example, Furnival and Wilson (1974). Similar algorithms for nonlinear time series analysis are worth developing.

As when using the ACE algorithm, care must be exercised in using the BRUTO algorithm in time series analysis. Strong serial correlation of the data might slow down the convergence of the algorithm. In fact, the convergence can be extremely slow in some cases. For illustration, consider the linear AR(3) model

$$y_t = .8y_{t-1} - .6y_{t-3} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are iid $N(0, 1)$. We generated 200 observations from the model and applied the BRUTO algorithm to the data with a maximum order of 5. The first iteration of the BRUTO algorithm selects a linear transformation of lag 4 (i.e., $y_{t-4}$), which is not in the model. Nevertheless, this selection is reasonable because the model has a maximum serial autocorrelation at lag 4. This can be easily seen by rewriting the model as

$$y_t = .64y_{t-2} - .96y_{t-4} + .36y_{t-6} + \varepsilon_t + .8\varepsilon_{t-1} - .6\varepsilon_{t-3},$$

for which $y_{t-4}$ has the highest coefficient. Unfortunately, such an erroneous selection causes problems in the subsequent selections because of the carry-over effect of $\hat{f}_4(y_{t-4})$. More specifically, the second step of the algorithm selects a linear transformation of $y_{t-1}$, which is a correct selection. But because the transformation of $y_{t-1}$ is based on the partial residual $y_t - \hat{f}_4(y_{t-4})$, it does not show the full effect of $y_{t-1}$ on $y_t$. This type of problem will not occur if $y_{t-1}$ and $y_{t-4}$ are uncorrelated. In the remaining iterations, the algorithm continuously attempts to improve the transformations of $y_{t-1}$ and $y_{t-3}$ and to reduce the effect of $y_{t-4}$. But such an improving process takes a long time to converge.

There are differences between the two proposed modeling procedures. For instance, unlike the best subset procedure, the procedure based on the BRUTO algorithm is automatic once the maximum order is given. Of course, both procedures have their own merits. For instance, properties of the ACE algorithm are much better understood than those of the BRUTO algorithm. Our limited experience indicates that neither algorithm is universally superior to the other. Thus we suggest that both algorithms be used in a real application, to provide a means for checking the identification of the unknown functions. If the two proposed modeling procedures suggest two different models, then care must be taken to select a final model. In this case we suggest that two parametric models be built based on the results of the procedures. A final model can then be chosen by using the conventional model selection techniques such as the AIC criterion, out-of-sample forecasting performance, and some parametric bootstrap procedures (Tsay 1992). Of course, in the event that the two competing models are nested, we can use the usual likelihood ratio test to select the final model.

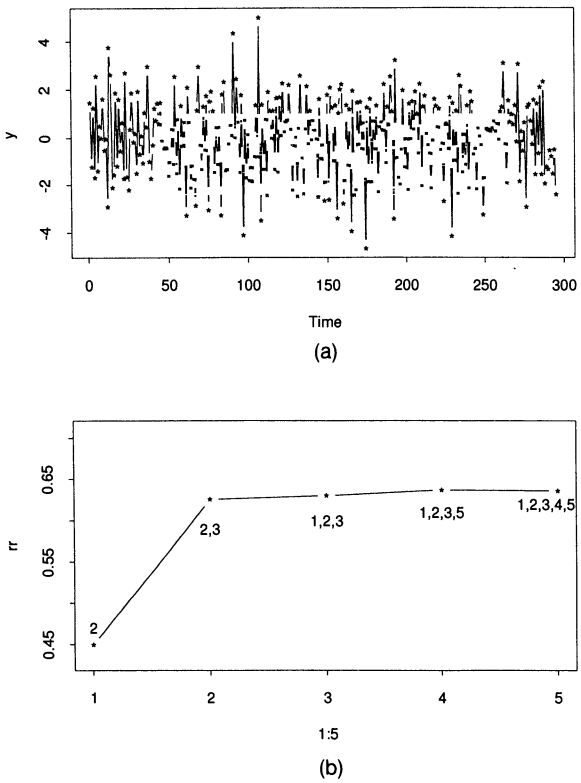


*Figure 1. (a) Time Plot of the Simulated Sine-Function AR Process of Example 1. (b) The $R^2$ of the Best Subset Autoregression Versus the Number of Lagged Variables Used for the Data Shown in (a).*

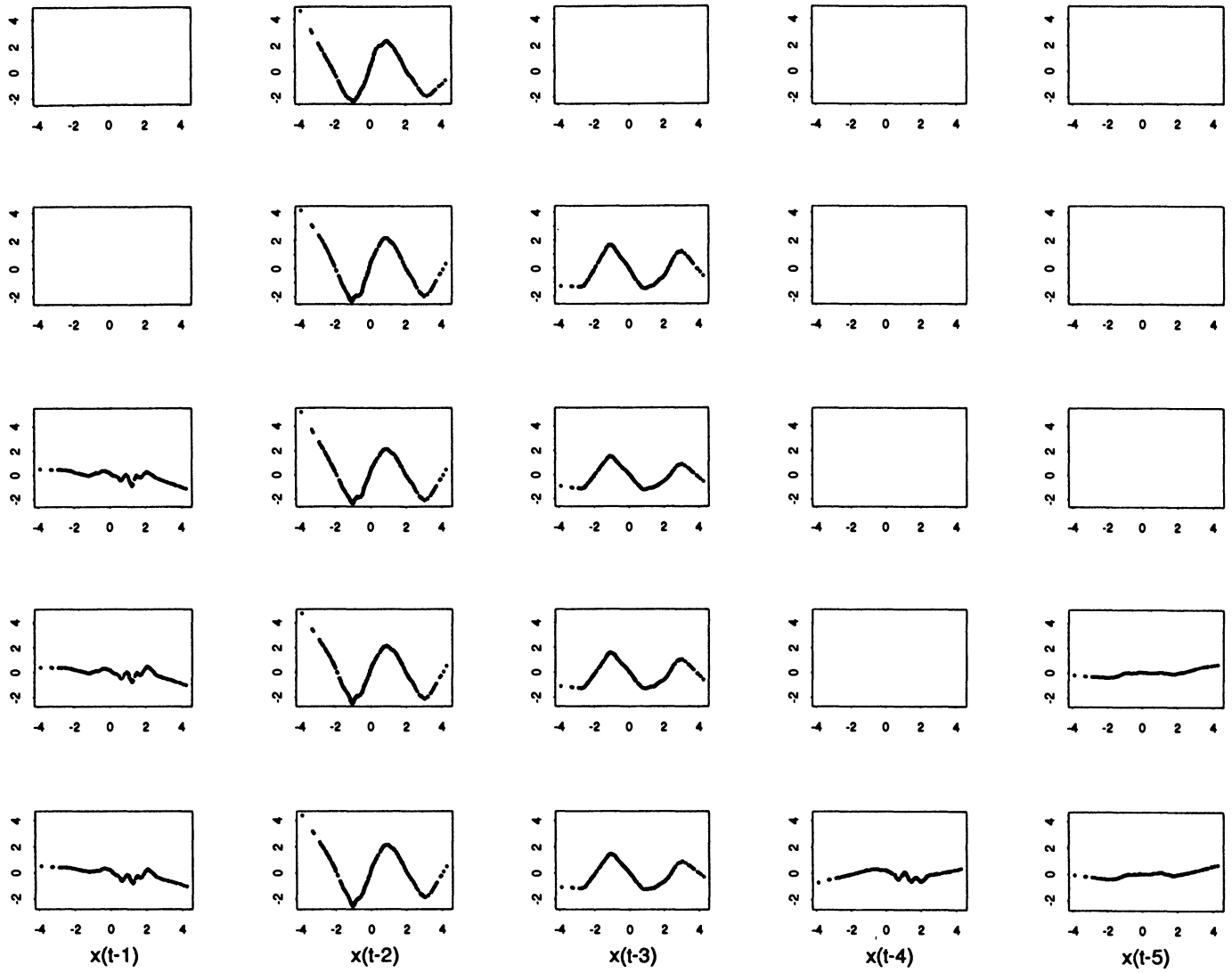


*Figure 2. Results of the ACE Algorithm for the Best Subset Autoregressions of Size i = 1, . . . , 5 for Example 1. The first row corresponds to the best subset autoregression of size 1, where the plot is on the second column indicating that $y_{t-2}$ is the explanatory variable to use. The rest of the plots are defined in a similar way.*

## 5. APPLICATIONS

We now apply the two proposed procedures to some real and simulated examples. In these applications, we use the following modeling process:

1. Select the lag indices $i_1, i_2, \ldots, i_p$ and $j_1, \ldots, j_q$ based on the results of the best subset procedure with ACE algorithm and of the BRUTO algorithm. If necessary, we entertain several tentative models for further investigation.
2. Specify the functional forms of a tentative model based on results of Step 1.
3. Estimate the postulated model by a conditional least squares method and check the fitted model. If necessary, go to Step 2 and refine the model.

### 5.1 Simulated Examples

We begin the illustration with two simulated examples of NAAR models. In the simulation the innovational series $\{\varepsilon_t\}$ are iid $N(0, 1)$.

*Example 1.* In this example we consider an AR process with sine functions. Figure 1 (a) shows 300 observations generated from the model

$$y_t = 1.5 \sin\left(\frac{\pi}{2} y_{t-2}\right) - 1.0 \sin\left(\frac{\pi}{2} y_{t-3}\right) + \varepsilon_t.$$

With a maximum order of 5, we first applied the best subset approach to the data. Figure 1 (b) shows the $R^2$ of the best subset regression versus the number of lagged variables used. As expected, the "true" model with lags (2, 3) stands out as a good candidate model. Figure 2 shows the results of the best subset regresssions of sizes $i = 1, \ldots, 5$. The first row is the best subset regression of size 1; the plots indicate that $y_{t-2}$ is the explanatory to use and that a sine function is appropriate. The second row gives the best subset regression of size 2; the plots show that $y_{t-2}$ and $y_{t-3}$ are the lagged variables and that trigonometric series are appropriate functionals. The remaining rows work in a similar way. Overall, the best subset modeling procedure via the ACE algorithm

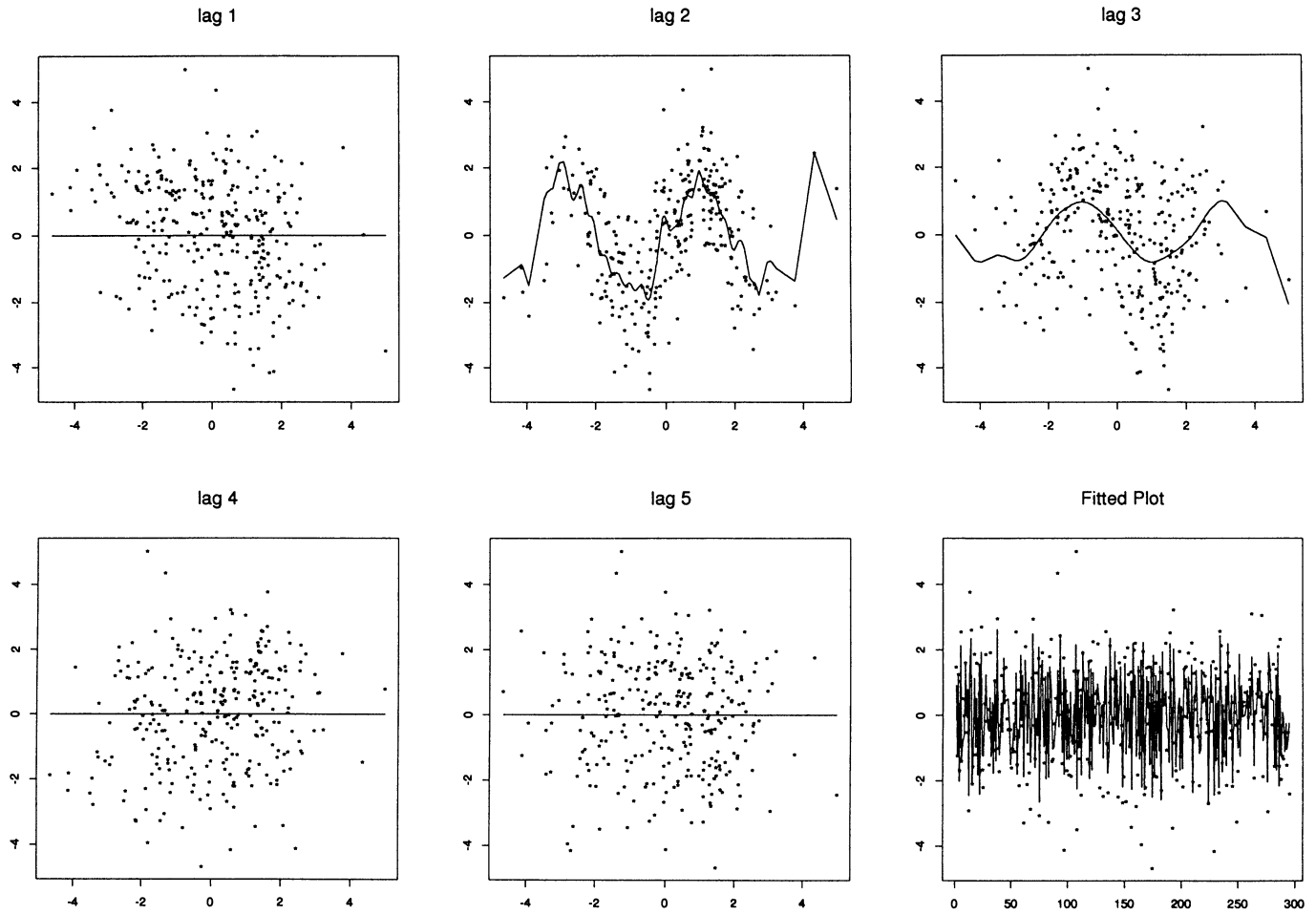


*Figure 3. Results of the BRUTO Algorithm for Example 1. The solid lines are the suggested transformations, and the points are the scatterplots of $y_t$ versus $y_{t-i}$ for $i = 1, \ldots, 5$. The plot on the lower right corner is the fitted value plot.*

appears to work well in this particular instance. It is reassuring to see that the functionals are close to 0 for overfitted lagged variables in rows 3–5.

Next, we applied the BRUTO algorithm to the data. With maximum order 5, the algorithm takes 11 iterations to converge and the transformations are on lags 2 and 3. The bandwidth for the transformation of lag 2 is .2623 and that of lag 3 is .8852. Figure 3 shows the output of the algorithm. The solid lines there are the suggested transformations, and the points are the scatterplots of $y_t$ versus $y_{t-i}$ for $i = 1, \ldots, 5$. Again, the method seems to work well. In particular it suggests that lags 1, 4, and 5 should be dropped from the model. The solid line in the last plot of the second row is the fitted values of the data via the BRUTO algorithm.

In summary, both the best subset modeling approach and the BRUTO algorithm suggest the model

$$y_t = f_1(y_{t-2}) + f_2(y_{t-3}) + \varepsilon_t$$

for the data, where $f_i(\cdot)$'s are trigonometric functions. The agreement between the two proposed procedures is encouraging.

Finally, we illustrate a procedure for checking the additivity assumption, using the idea of Tukey's 1 degree of freedom test (see method (ii) in sec. 9.5.1 of Hastie and Tibshirani 1990, p. 264). This procedure is done basically to check that there are no interactions between the explanatory variables. Denote the estimates of $f_i(\cdot)$ by $\hat{f}_i(\cdot)$ and denote the residual by $\hat{\varepsilon}_t = y_t - \hat{f}_1(y_{t-2}) - \hat{f}_2(y_{t-3})$. To check additivity, we consider the linear regression

$$\hat{\varepsilon}_t = \beta_0 + \beta_1 \hat{f}_1(y_{t-2})\hat{f}_2(y_{t-3}) + e_t$$

and obtain the test statistic $nR^2$, where $n$ is the sample size and $R^2$ is the usual coefficient of determination of the preceding linear regression. Under the null hypothesis of no interaction between $y_{t-2}$ and $y_{t-3}$ and normality, the test statistic asymptotically follows $\chi^2$ distribution with 1 degree of freedom. For this particular example the test statistic is 3.34, which, as expected, is statistically insignificant at the usual 5% level.

*Example 2.* In this example we generated 300 observations from the model

$$y_t = .8 \log(1 + 3y_{t-1}^2) - .6 \log(1 + 3y_{t-3}^2) + \varepsilon_t.$$

Figure 4(a) shows the time plot of the data. Figure 4(b) is the $R^2$ of the best subset regression versus the number of

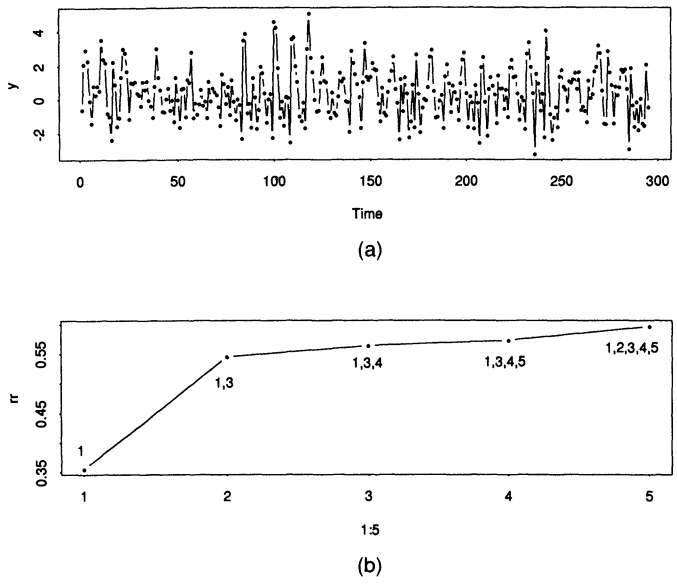


*Figure 4. (a) Time Plot of the Simulated Log-Function AR Process of Example 2. (b) The $R^2$ of the Best Subset Autoregression Versus the Number of Lagged Variables Used for the Data Shown in (a).*

lagged variables used when the maximum order is 5. Again, the "true" combination (1, 3) is easily identified as a tentative model. Figure 5 shows the functions of the best subset regressions of sizes 1 to 5. The result is, again, informative. But it is not easy to specify from the plots that $\log(1 + 3y_{t-i}^2)$ is the function to use, because there are other functions with similar shape. This is a common difficulty in using nonparametric techniques to specify the functional form of a nonlinear time series model.

For this second example the BRUTO algorithm only took two iterations to converge. It also correctly selected lags 1 and 3 as significant explanatory variables. The bandwidth for the transformation of lag 1 is 1.037 and that of lag 3 is 1.196. The result of the algorithm is shown in Figure 6. Again, the two proposed procedures work well in revealing the nonlinear characteristic of the model.

## 5.2 A Real Example

For an application of NAAR$X$ models, we consider some riverflow data of River Jökulsá Eystri of Iceland. The data consist of daily riverflow ($y_t$), precipitation ($z_t$), and tem-

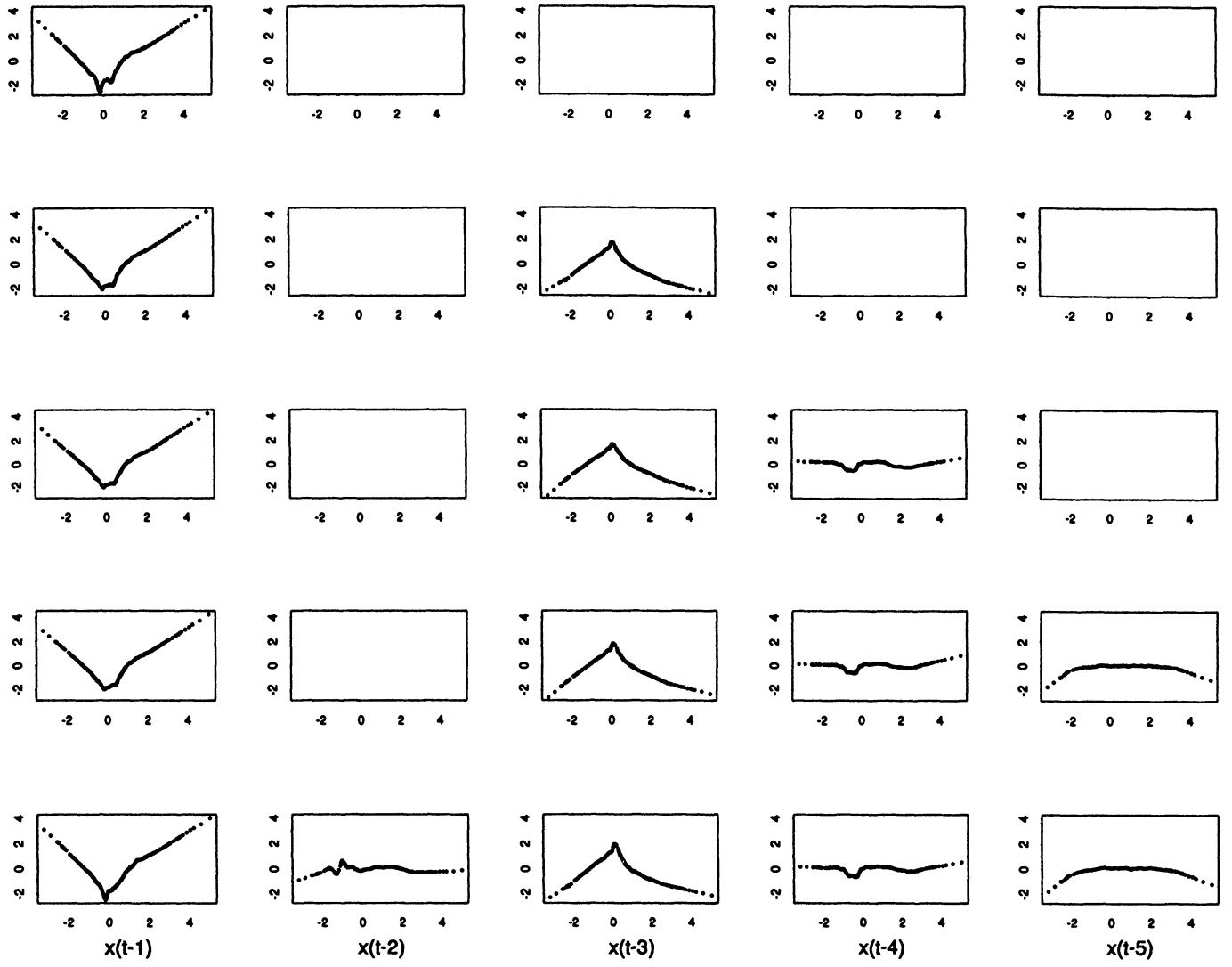


*Figure 5. Results of the ACE Algorithm for the Best Subset Regressions of Size $i = 1, \ldots, 5$ for Example 2. The plot can be read in the same way as Figure 2.*

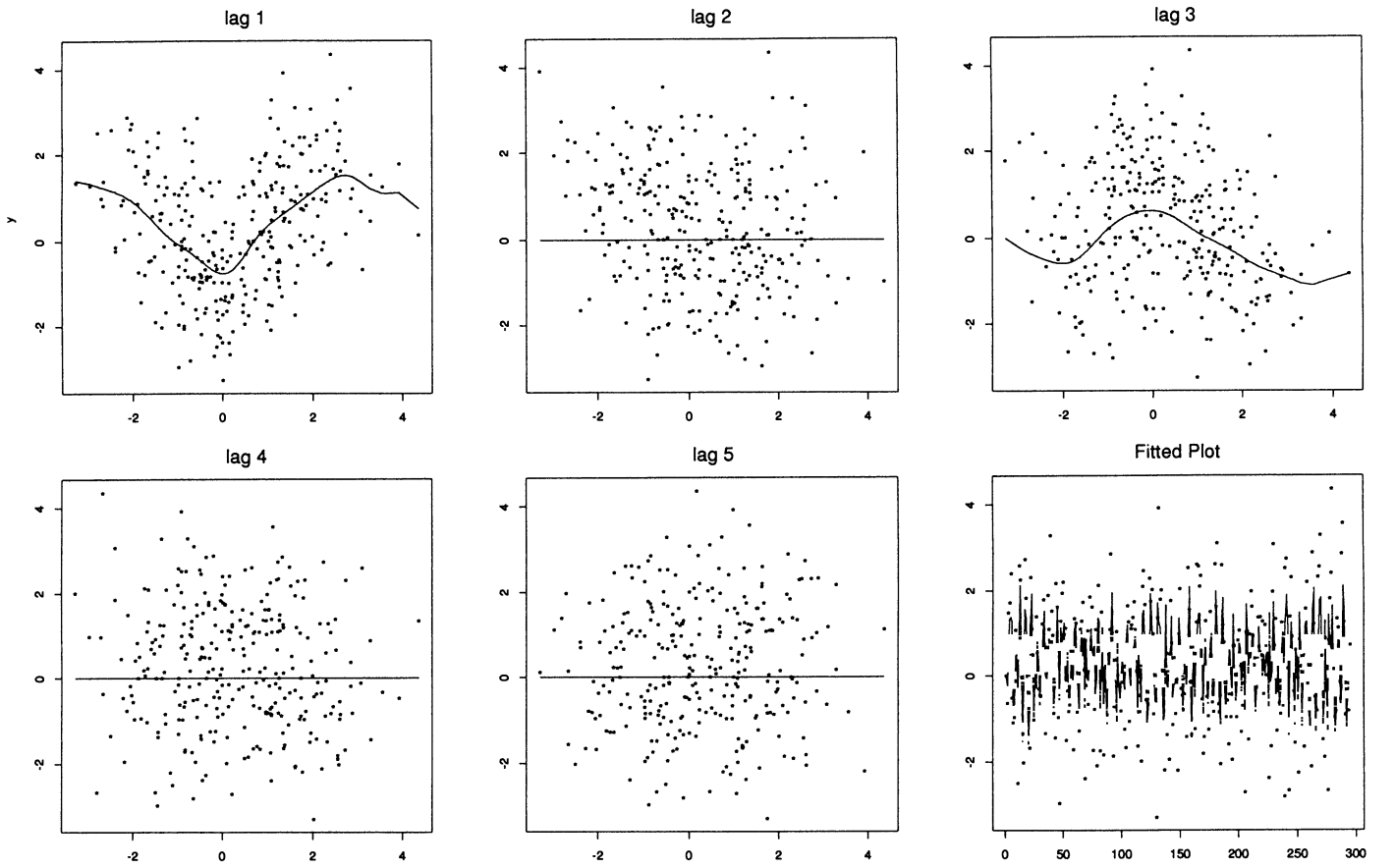


*Figure 6. Result of the BRUTO Algorithm for the Process of Example 2. The solid lines are the suggested transformations, and the points are the scatterplots of $y_t$ versus $y_{t-i}$ for $i = 1, \ldots, 5$. The plot on the lower right corner is the fitted value plot.*

perature ($x_t$) from January 1, 1972, to December 31, 1974. There are 1,096 observations. An important hydrological feature of this river is that there is a glacier on the drainage area. Consequently, temperature has certain influence on riverflow besides melting the snow. For further information of the data, see Tong (1990). The data were analyzed earlier by Tong, Thanoon, and Gudmundsson (1985) and Tong (1990), who used threshold autoregressive models. Figures 7(a)–(c) show the riverflow ($m^3/s$), temperature (°C), and precipitation (mm/day).

Using exogenous variables $x_t$ and $z_t$, we applied the best subset modeling approach to the data. The maximum order entertained for the lagged endogenous variables $y_{t-i}$ is 4, that for precipitation $z_{t-j}$ is 3, and that for temperature $x_{t-k}$ is 4. These maximum orders were selected by examining the serial correlations of $y_t$ and the cross-correlations between $y_t$ and $z_{t-j}$, $x_{t-k}$. Figure 8(a) shows the results of the best subset regressions of sizes 1 to 6. The size here denotes the number of explanatory variables used in the model. Figure 9(a) plots the $R^2$ of the best subset regression versus size. From the plot and for simplicity, it appears that $\{y_{t-1}, y_{t-2}, z_t, z_{t-1}, x_{t-1}, x_{t-3}\}$ could be a set of appropriate explanatory variables to use. In what follows, we entertained such a model.

Figure 8(b) is an enlarged version of the bottom row of Figure 8(a). An examination of the transformations shown there suggests that an NAAR$X$ model with simple piecewise linear functionals might be sufficient in describing the data. Consequently, we postulated a parametric model with piecewise linear coefficients. In model checking, we added two lagged endogenous variables $y_{t-3}$ and $y_{t-4}$, resulting in the model suggested by the best subset regression of size 8; see Figure 9(a). Consequently, the NAAR$X$ model entertained is

$$\begin{aligned} y_t = {} & c + \phi_{1,1} y_{t-1} + \phi_{1,2} y_{t-1} I(y_{t-1} \geq c_1) \\ & + \phi_{1,3} y_{t-1} I(y_{t-1} \geq c_2) + \phi_2 y_{t-2} + \phi_3 y_{t-3} \\ & + \phi_4 y_{t-4} + \beta_1 z_t + \beta_2 z_{t-1} + \omega_{1,1} x_{t-1} \\ & + \omega_{1,2} x_{t-1} I(x_{t-1} \geq c_3) + \omega_{3,1} x_{t-3} \\ & + \omega_{3,2} x_{t-3} I(x_{t-3} \geq c_4) + \varepsilon_t, \end{aligned} \tag{6}$$

where $c_1 = 27$, $c_2 = 100$, $c_3 = 1$, $c_4 = 1$, and the coefficients are

| Par. | $c$ | $\phi_{1,1}$ | $\phi_{1,2}$ | $\phi_{1,3}$ | $\phi_2$ | $\phi_3$ | $\phi_4$ | $\beta_1$ | $\beta_2$ | $\omega_{1,1}$ | $\omega_{1,2}$ | $\omega_{3,1}$ | $\omega_{3,2}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Est. | 1.88 | 1.15 | .04 | −.08 | −.31 | .19 | −.12 | .36 | −.20 | −.07 | 1.67 | .05 | −1.39 |
| Std. | .92 | .04 | .02 | .01 | .04 | .04 | .03 | .03 | .03 | .06 | .14 | .06 | .14 |

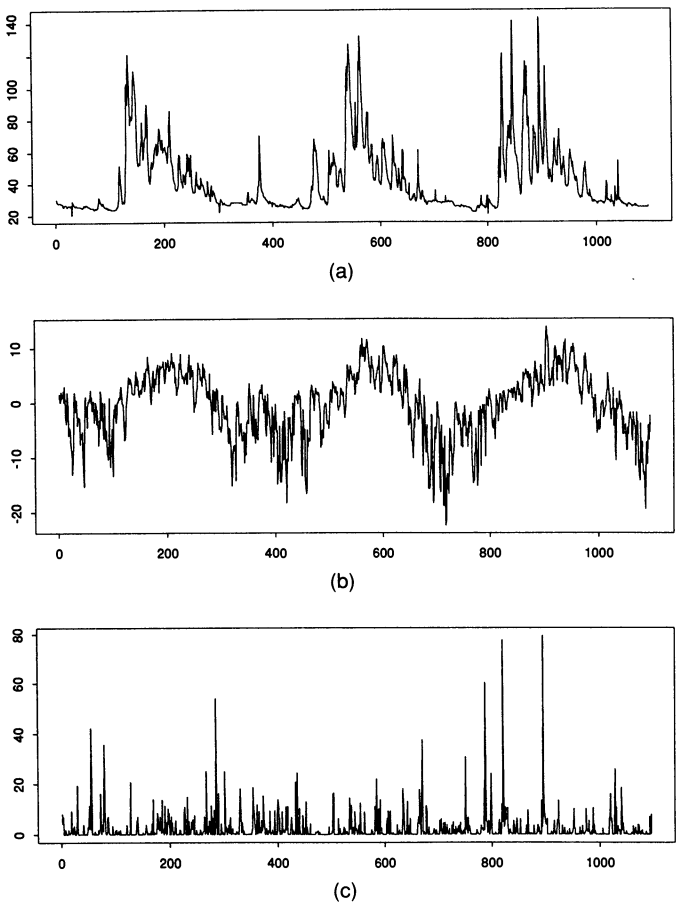


*Figure 7. Time Plots of River Jökulsá Eystri Riverflow Data. (a) Daily riverflow $y_t$ ($M^3/s$). X-axis is the time index; Y-axis is riverflow. (b) Daily temperature $x_t$ (°C). X-axis is the time; Y-axis is temperature. (c) Daily precipitation $z_t$ (mm/day). X-axis is the time; Y-axis is precipitation.*

The structure parameters $c_1$, $c_2$, $c_3$, and $c_4$ are obtained by minimizing the sum of squares of residuals, and the estimated parameters are obtained using conditional least squares estimation. From Figure 8(b) it is seen that the estimates of these structure parameters are close to those suggested by the ACE algorithm. The residual variance of Model (6) is 33.15, which corresponds to an $R^2$ of 92.48%. The fitted values and residuals are shown in Figure 9(b) and 9(c). The residual plot shows some isolated spikes, indicating that, as expected, normality is not adequate in describing the riverflow data. There are, however, no significant serial correlations in the residuals; only lag 7 has a marginal correlation of −.10. Overall, the nonlinear model (6) appears to fit the data reasonably well.

We also applied the variable selection procedure to the data. With the same maximum orders as before, the BRUTO algorithm took six iterations to converge and selected the explanatory variables $\{y_{t-1}, y_{t-2}, z_t, x_t\}$ where, except for $y_{t-2}$, all the functions are nonlinear. Based on this selection and some further refinement, we arrived at the model

$$\begin{aligned} y_t = {} & 1.52 + 1.30y_{t-1} - .01y_{t-1}I(y_{t-1} \geq 24) \\ & - .09y_{t-1}I(y_{t-1} \geq 100) - .46y_{t-2} + .22y_{t-3} \\ & - .10y_{t-4} + .27z_t + .19z_tI(z_t \geq 35) - .19z_{t-1} \\ & + .12x_t + .50x_tI(x_t \geq -2) - .05x_{t-1} + e_t. \end{aligned} \quad (7)$$

The residual variance of model (7) is 35.32, which is larger than that of Model (6). Because these two NAAR$X$ models use the same number of parameters, we select Model (6) by using the AIC criterion. This selection is further supported by the mean squared errors of out-of-sample forecasts. Using the same forecasting procedure as that mentioned in the following discussion, we found that Model (6) outperforms Model (7) in all but one of the forecasts considered.

*Discussion.* The temperature effect on riverflow shown in the NAAR$X$ Model (6) is interesting. When the temperature is below 1°C, riverflow is not significantly affected. On the other hand, when the temperature is higher than 1°C, the riverflow of the next day increases substantially, presumably is due to snow melting. The reverse influence of temperature 3 days earlier on riverflow is understandable, because the effect of snow melting cannot last for long. Of course, one needs to keep in mind the dynamic nature of $y_t$ in interpreting the influence of temperature 3 days earlier.

The piecewise function of $y_{t-1}$ in (6) indicates that the daily riverflow $y_t$, after adjusting the effects of precipitation and temperature, has different dynamic properties depending on the level of the previous riverflow $y_{t-1}$. This appears to be reasonable.

To appreciate the contribution of the NAAR$X$ Model (6), we compare it with Tong's threshold model (1990) and a linear transfer function model. Tong (1990) used temperature as the threshold variable and fitted the following model:

$$\begin{aligned} y_t = {} & c_1 + (1, \ldots, 6)y_t + (0, \ldots, 5)z_t \\ & + (0, \ldots, 3)x_t + e_{1t} \quad \text{if } x_t \leq -2 \\ = {} & c_2 + (1, \ldots, 8)y_t + (0, \ldots, 7)z_t \\ & + (0, \ldots, 7)x_t + e_{2t} \quad \text{if } x_t > -2, \end{aligned}$$

where $(1, \ldots, 6)y_t$ means that $y_{t-1}, y_{t-2}, \ldots, y_{t-6}$ are included in the model and the other terms are defined in a similar fashion. The overall residual variance of this threshold model is 31.77, which is slightly smaller than that of our NAAR$X$ model. Using linear transfer function analysis, we obtained the model

$$\begin{aligned} y_t = {} & 41.03 + (.35 + .20B)z_t + \frac{.27 + .21B + .18B^2}{1 - .71B}x_t \\ & + \frac{1 - .12B^7}{1 - 1.17B + .42B^2 - .22B^3 + .07B^4}a_t, \end{aligned} \quad (8)$$

where $B$ is the usual backshift operator such that $Bz_t = z_{t-1}$ and the residual variance of $a_t$ is 38.57. In Model (8), $t$ ratios of all the parameters are greater than 2.0 in modulus. The residual series of this linear model also has no significant serial correlations except for lag 4, which is −.07.

In terms of residual variance, Tong's threshold model appears to be best. But the NAAR$X$ Model (6) has fewer parameters and is the preferred model based on the AIC criterion. Further, the NAAR$X$ model produces better out-of-sample forecasts. For the three models entertained, we reestimated the parameters using only data of the first 2 years and performed naive out-of-sample forecasts of Steps 1–12 for the entire third year. The real values of the exogenous

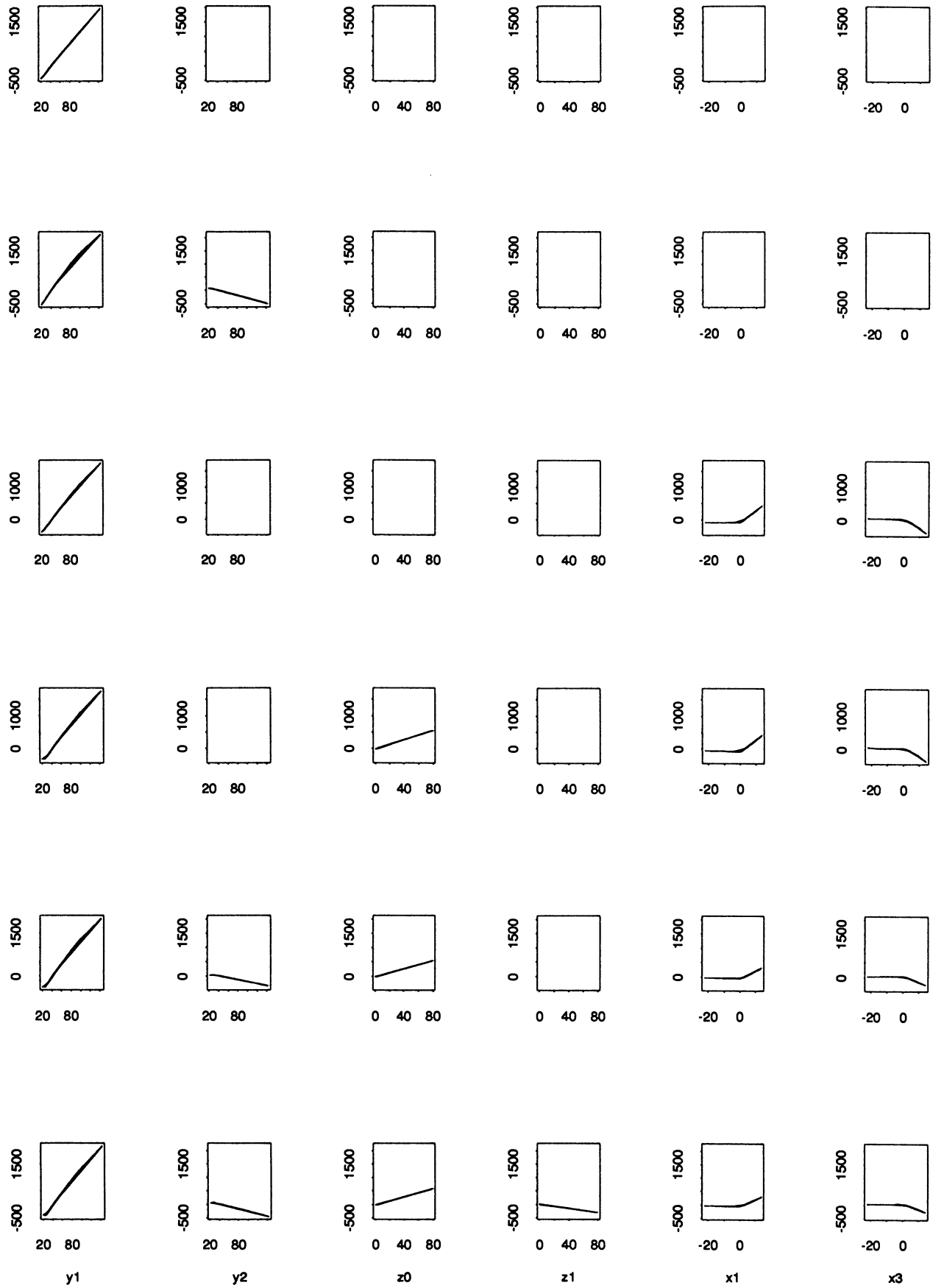


*Figure 8a. Results of the ACE Algorithm for the Best Subset Regressions of Size $i = 1, \ldots, 6$ for the Riverflow Data Example. The plot can be read in the same way as Figure 2. The first column is for variable $y_{t-1}$, the second column for $y_{t-2}$, the third column for $z_t$, the fourth column for $z_{t-1}$, the fifth column for $x_{t-1}$ and the sixth column for $x_{t-3}$, where $y_t$ indicates the riverflow series, $z_t$ indicates the precipitation, and $x_t$ indicates the temperature.*

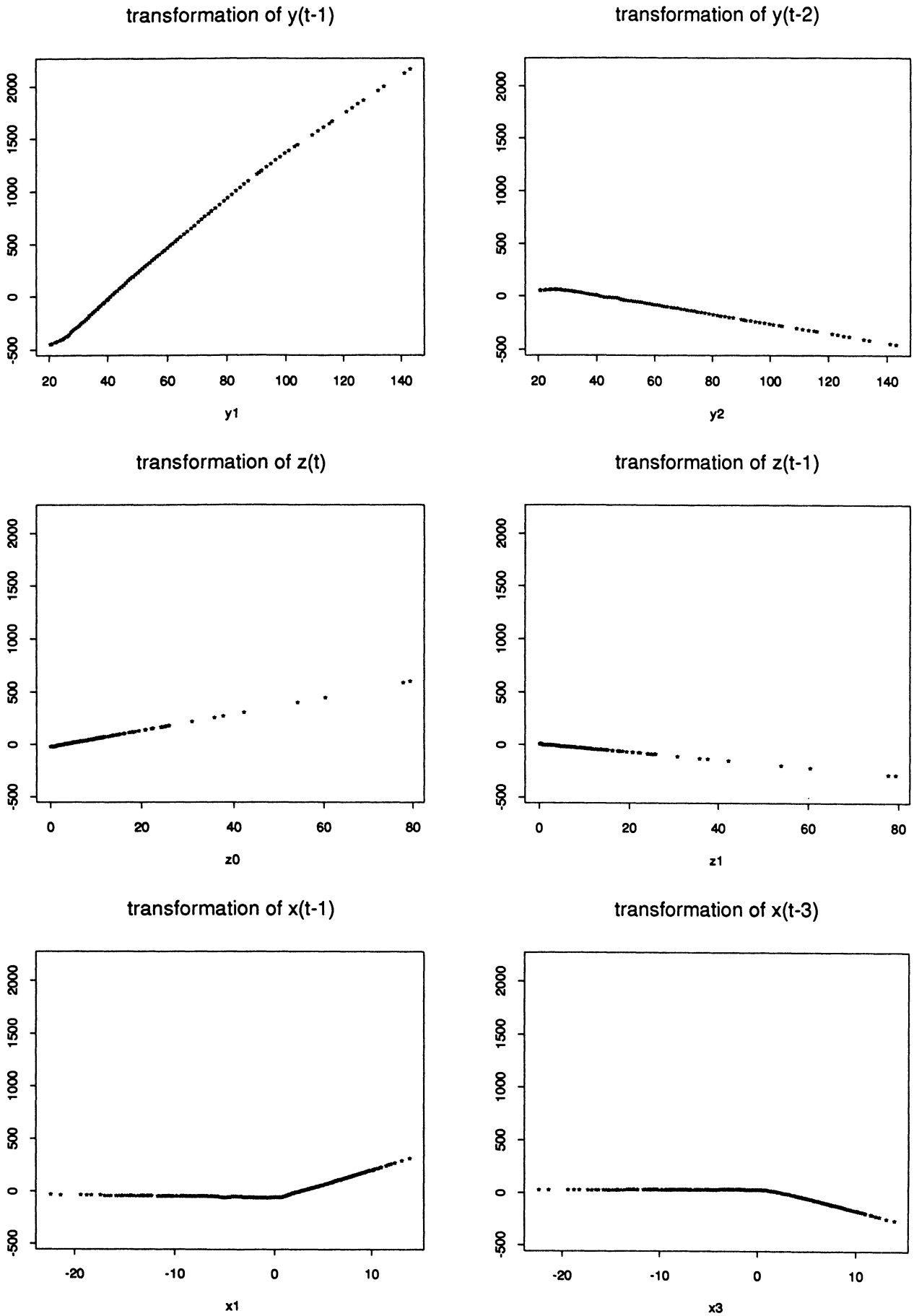


*Figure 8b. An Enlarged Version of the Bottom Row of Figure 8.*

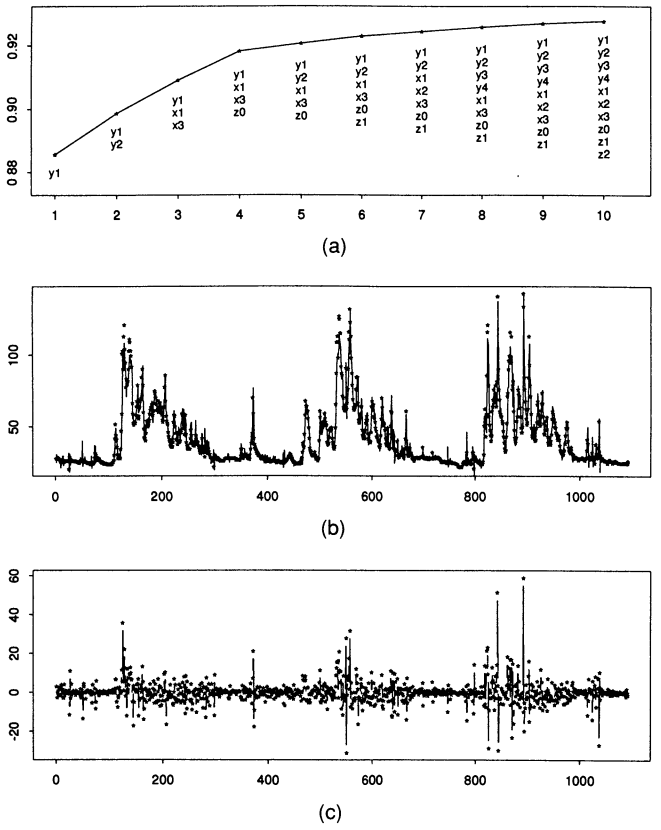


*Figure 9. (a) The $R^2$ of the Best Subset Regressions Versus the Number of Explanatory Variables Used for the Riverflow Example. X-axis is the size of explanatory variable used; Y-axis is the $R^2$. (b) Fitted Plot. The solid line is the fitted value of the river flow series and the points are the scatterplot of the river flow series versus time index. (c) Residuals of the Fitted NAARX Model.*

variables were used in the forecast. We then computed the mean squared errors of forecasts of the three models. The results are given below:

| Lead time | NAAR$X$ | TAR | Linear |
|---|---|---|---|
| 1 | 65.52 | 66.67 | 81.99 |
| 2 | 142.01 | 153.39 | 221.60 |
| 3 | 175.88 | 196.37 | 335.01 |
| 4 | 203.03 | 231.85 | 376.20 |
| 5 | 230.02 | 258.92 | 418.91 |
| 6 | 256.89 | 283.16 | 472.71 |
| 7 | 279.32 | 300.27 | 523.92 |
| 8 | 295.25 | 309.88 | 566.12 |
| 9 | 308.09 | 318.83 | 613.41 |
| 10 | 315.07 | 323.68 | 661.13 |
| 11 | 320.56 | 335.34 | 725.82 |
| 12 | 321.56 | 350.41 | 801.29 |

From the table, it is clear that the linear model performs poorly in out-of-sample forecasts, and that the NAAR$X$ model consistently produces the best forecasts among the three models considered.

Finally, we agree that it is hard to justify that riverflow follows an additive model. Indeed, our additivity checking shows certain significant interactions between the explanatory variables when 5% critical values of chi-squared distributions are used. But because the normality assumption is clearly violated, we do not know what effect such a violation has on the critical values. This is an issue that needs further investigation. On the other hand, we believe that the additive model used provides a better approximation to the behavior of the data, because it performs well in out-of-sample forecasts and has high $R^2$ value.

## 6. SUMMARY

In this article we considered a class of nonlinear additive autoregressive models with exogenous variables for nonlinear time series analysis and used two backfitting procedures to specify such models. The model specification was achieved by using the methods of best subset regression and variable selection in regression analysis. The modeling procedures proposed were shown to be useful in analyzing simulated series and a real example. The results suggest that nonlinear additive models in conjunction with nonparametric techniques could be useful in nonlinear time series analysis.

[*Received September 1991. Revised October 1992.*]

## REFERENCES

Akaike, H. (1974), "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control,* AC-19, 716–722.

Auestad, B., and Tjøstheim, D. (1990), "Identification of Nonlinear Time Series: First Order Characterization and Order Determination," *Biometrika,* 77, 669–687.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association,* 80, 580–619.

Chen, R. (1990), "Two Classes of Nonlinear Time Series Models," unpublished Ph.D. dissertation, Carnegie Mellon University, Dept. of Statistics.

Chen, R., and Tsay, R. S. (1993), "Functional-Coefficient Autoregressive Models," *Journal of the American Statistical Association,* 88, 298–308.

Friedman, L. H. (1991), "Multivariate Adaptive Regression Splines," (with discussion), *The Annals of Statistics,* 19, 1–141.

Friedman, L. H., and Stuetzle, W. (1982), "Smoothing of Scatterplots," technical report, Stanford University, Dept. of Statistics.

Furnival, G., and Wilson, R. (1975), "Regression by Leaps and Bounds," *Technometrics,* 16, 499–511.

Haggan, V., and Ozaki, T. (1981), "Modeling Nonlinear Vibrations Using an Amplitude-Dependent Autoregressive Time Series Model," *Biometrika,* 68, 189–196.

Hart, J. D., and Vieu, P. (1990), "Data-Driven Bandwidth Choice for Density Estimation Based on Dependent Data," *The Annals of Statistics,* 18, 873–890.

Hastie, T. J. (1989), Discussion of "Flexible Parsimonious Smoothing and Additive Modeling" by J. Friedman and B. Silverman, *Technometrics,* 31, 23–29.

Hastie, T. J., and Tibshirani, R. J. (1991), *Generalized Additive Models,* London: Chapman and Hall.

Jones, D. A. (1978), "Nonlinear Autoregressive Processes," *Proceedings of the Royal Statistical Society,* Ser. A, 360, 71–95.

Lewis, P. A. W., and Stevens, J. G. (1991), "Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS)," *Journal of the American Statistical Association,* 86, 864–877.

Koyak, R. (1990), "Consistency for ACE-Type Methods," *The Annals of Statistics,* 18, 742–757.

Robinson, P. M. (1983), "Nonparametric Estimation for Time Series Models," *Journal of Time Series Analysis,* 4, 185–208.

Tong, H. (1978), "On a Threshold Model," in *Pattern Recognition and Signal Processing,* ed. C. H. Chen, Amsterdam: Sijthoff and Noordhoff.

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

Tong, H., Thanoon, B., and Gudmundsson, G. (1985), "Threshold Time Series Modeling of Two Icelanic Riverflow Systems," in *Time Series Analysis in Water Resources*, ed. K. W. Hipel, American Water Research Association.

Tsay, R. S. (1992), "Model Checking via Parametric Bootstraps in Time Series Analysis," *Applied Statistics,* 41, 1–15.