Statistica Sinica **7**(1997), 451-472

A UNIFIED APPROACH TO ESTIMATING AND MODELING LINEAR AND NONLINEAR TIME SERIES

Cathy W. S. Chen*, Robert E. McCulloch and Ruey S. Tsay

**Feng-Chia University and University of Chicago*

Abstract: In this article, we propose a unified approach to estimating and modeling univariate time series. The approach applies to both linear and nonlinear time series models and can be used to discriminate non-nested nonlinear models. For example, it can discriminate between threshold autoregressive and bilinear models or between autoregressive and moving average models. It can also be used to estimate and discriminate between symmetric and asymmetric conditional heteroscedastic models commonly used in volatility studies of financial time series. The proposed approach is based on Gibbs sampling and may require substantial amounts of computing in some applications. We illustrate the proposed approach by some simulated and real examples. Comparison with other existing methods is also discussed.

Key words and phrases: Bayesian model selection, bilinear model, Gibbs sampler, mixed model, stochastic volatility, threshold autoregressive model.

1. Introduction

Estimation and model selection are two main components of time series analysis. There are many results available in the literature that concern parameter estimation of a given model and model selection within a specified class of models. For example, the exact maximum likelihood method was widely investigated in 1980s for autoregressive moving-average (ARMA) models, e.g. Ansley (1979), Jones (1980), Hillmer and Tiao (1979) among others, and the conditional least squares approach was proposed for nonlinear models, e.g. Tong (1990) and the references therein. For model selection, some popular model selection criteria such as AIC and its variants are commonly used to select the order of an autoregressive process or a threshold autoregressive models. See, for instance, Priestley (1981), Brockwell and Davis (1991) and Tong (1990).

There is, however, no unified approach or program that can be used to estimate most of the linear and nonlinear models considered in the literature. For example, special packages are needed to apply bilinear models, threshold models or Markov switching models. Furthermore, there is little discussion of model selection across different classes of nonlinear models. Much work on model selection in the literature focuses on nested models for which the traditional maximum

likelihood ratio tests or Largange multiplier tests or information criterion functions apply. For non-nested models, model discrimination becomes much more involved, especially when the competing models are nonlinear. In the time series literature, Li (1993) adopts the idea of separate families of hypotheses of Cox (1962) and proposes a test statistic for discriminating bilinear and threshold models. The test statistic has an asymptotic chi-squared distribution with one degree of freedom. However, Li's test is closely related to the method of selecting a model with smaller residual variance and is not applicable to other nonlinear models.

The purpose of this paper is to propose a unified approach that can be used to estimate most of the univariate time series models available in the literature and to select an appropriate model for a time series when the candidate models may be non-nested nonlinear. More specifically, our objective is to consider an approach that is widely applicable in univariate time series analysis. The models can be linear or nonlinear, and the approach can discriminate between non-nested nonlinear models. The proposed approach is based on Gibbs sampling and requires some prior specification. In particular our approach to model selection allows each observation to select one of the candidate models. The key prior specification here is the probability that an individual observation is generated by a specified model given that both observations adjacent in time are generated by that same model. (See equation (5) below.) Sensitivity analysis of prior specification will be discussed later.

Because the proposed approach uses Gibbs sampling, it may require substantial computing time in some applications. Our goal is not to develop the most efficient approach for univariate time series analysis, but a unified approach that is applicable to most parametric models. In a given application where the entertained models are specified, it is often possible to reduce the computing time by some special algorithm or theoretical derivation. However we shall not focus on those special issues.

The paper is organized as follows. In Section 2, we give the general framework of models considered in this paper and show that many time series models considered in the literature are special cases of our model. Section 3 considers model estimation via the Gibbs sampling. In particular, we treat starting values of the time series and the innovational series as parameters and consider the conditional likelihood function of a parameter given the others. We also discuss methods for implementing the Gibbs sampler when the parameter under study is nonlinear, e.g. the moving-average parameters in an ARMA model. Section 4 is devoted to model discrimination. Here a simple switching framework is used in which the competing non-nested nonlinear models become submodels of a mixture. Under this framework, each individual observation can select its own model

from the mixture. The posterior probability that particular observations are associated with a particular model can then be used to select an appropriate model for the whole series. Advantages of the proposed method of model selection are discussed. While this idea for model discrimination was used in McCulloch and Tsay (1994) and in George, McCulloch and Tsay (1996) for some special models, we implement it differently in this paper and extend it to a procedure for general non-nested nonlinear models. Finally, we illustrate the proposed approach by some simulated and real examples in Section 5.

2. A General Nonlinear Model

The model considered in this paper is

$$\begin{aligned} y_t &= f(y_{t-1}, \ldots, y_{t-p}; a_{t-1}, \ldots, a_{t-q}; \boldsymbol{\beta}_f) + a_t \\ a_t &= g_t \epsilon_t \\ g_t &= g(y_{t-1}, \ldots, y_{t-u}; a_{t-1}, \ldots, a_{t-v}; g_{t-1}, \ldots, g_{t-w}; \boldsymbol{\beta}_g), \end{aligned} \tag{1}$$

where y_t is a univariate time series, $f(\cdot)$ and $g(\cdot)$ are two known functions with finite-dimensional parameter vectors $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_g$, respectively, p, q, u, v, and w are non-negative integers, and $\{\epsilon_i\}$ is a sequence of independent and identically distributed random variables with mean zero and variance one. The function $g(\cdot)$ is assumed to be positive; it governs the evolution of the volatility of the innovational series a_t. For simplicity, we focus on the case that the ϵ_t's are standard normal random variables, i.e. a_t is conditionally normal. However, it is easily seen that ϵ_t can be any continuous random variables with a well defined density function.

Model (1) is a general model, because it encompasses many commonly used models in the literature. Some specific examples are:

1. If $g(\cdot) = \beta_1$, which is a positive constant, and $f(\cdot) = \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i}$; then model (1) reduces to the well-known ARMA of Box, Jenkins and Reinsel (1994).
2. If $f(\cdot) = 0$ and $g^2(\cdot) = \gamma_0 + \sum_{i=1}^{q} \gamma_i a_{t-i}^2$, where $\gamma_0 > 0$ and $\gamma_i \geq 0$, then the model becomes the well-known conditional autoregressive heteroscedastic (ARCH) model of Engle (1982). The ARCH model and its variants are widely used in finance to model the volatility of a security return.
3. If $f(\cdot) = 0$ and $g^2(\cdot) = \gamma_0 + \sum_{i=1}^{v} \gamma_i a_{t-i}^2 + \sum_{i=1}^{w} \lambda_i g_{t-i}^2$, where $\gamma_0 > 0$, $\gamma_i \geq 0$ and $\lambda_i \geq 0$, then we have the generalized ARCH (GARCH) model of Bollerslev (1986).
4. If $f(\cdot) = 0$ and $g(\cdot) = \exp(\gamma_0 + \sum_{i=1}^{u} \beta_i y_{t-i} + \sum_{j=1}^{v} \gamma_j a_{t-j})$, then model (1) becomes a stochastic volatility model in which the conditional variance of the series is related to past observations and past innovations. This model is

similar to that in Tsay (1987) and can be extended to include models that allow for asymmetric responses to positive and negative innovations.

5. If $g(\cdot) = \beta_1 > 0$, a constant, and $f(\cdot) = \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i} + \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij} y_{t-i} a_{t-j}$, then model (1) becomes the bilinear model of Granger and Andersen (1978) and Subba Rao (1981).
6. If $f(\cdot) = \phi_0^{(i)} + \sum_{j=1}^{p} \phi_j^{(i)} y_{t-i}$ and $g(\cdot) = \sigma^{(i)} > 0$ for $r_{i-1} \leq y_{t-d} \leq r_i$, where d is a positive integer and r_i's are real numbers satisfying $-\infty = r_0 < r_1 < \cdots < r_k = \infty$, then model (1) becomes the threshold autoregressive (TAR) model of Tong (1978, 1990).

Model (1) also provides a framework to combine different time series models. For example, if $f(\cdot) = \phi_0 + \phi_1 y_{t-1} - \theta_1 a_{t-1}$ and $g(\cdot) = \omega_0 + \omega_1 a_{t-1} > 0$ almost surely, then y_t is an ARMA process with a concurrent bilinear innovation. Such an innovational series also shows stochastic volatility as that of an ARCH model. In Section 4, we use model (1) to develop a switching model for discrimination of non-nested nonlinear models.

3. Estimation

In this section, we discuss a general approach to parameter estimation of model (1). The proposed approach is Bayesian and makes use of Gibbs sampling. In particular, we assume the time series y_t starts at time $t = 1$ with unknown starting values, lagged innovations, and lagged g values. We treat these initial values as unknown parameters of the model and estimate them jointly with other parameters. This marks a "big" difference between the proposed approach and many existing estimation methods, because those existing methods assume either the starting values are zero or the process under study is stationary. (See, for example, Brockwell and Davis (1991).) The idea of treating starting values and innovations of a time series process as unknown parameters has been used previously in the literature primarily for ARMA models. For example, in exact likelihood estimation of ARMA models, those starting values and innovations are estimated by using the dynamic structure of the data. However, for nonlinear models no formal study is currently available.

Consider model (1). Let $p^* = \max\{p, u\}$, $q^* = \max\{q, v\}$, $\boldsymbol{y}_0 = (y_{-p^*+1}, y_{-p^*+2}, \ldots, y_0)'$ be the starting values of y_t, $\boldsymbol{a}_0 = (a_{-q^*+1}, a_{-q^*+2}, \ldots, a_0)'$ be the starting innovations, and $\boldsymbol{g}_0 = (g_{-w+1}, \ldots, g_0)'$ the starting lagged g values. Finally, let $\boldsymbol{\Omega} = (\boldsymbol{y}_0', \boldsymbol{a}_0', \boldsymbol{g}_0', \boldsymbol{\beta}_f', \boldsymbol{\beta}_g')'$ be the set of all parameters of model (1). For n observations $\{y_t\}_{t=1}^n$, let $\boldsymbol{Y}_t = (y_1, \ldots, y_t)'$. It is easily seen that the conditional mean and variance of y_t given $\boldsymbol{Y}_{t-1}$ and $\boldsymbol{\Omega}$ are

$$E(y_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\Omega}) = f(y_{t-1}, \ldots, y_{t-p}; a_{t-1}, \ldots, a_{t-q}) \equiv f_t$$
$$\text{var}(y_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\Omega}) = g^2(y_{t-1}, \ldots, y_{t-u}; a_{t-1}, \ldots, a_{t-v}; g_{t-1}, \ldots, g_{t-w}) \equiv g_t^2.$$

Therefore, the log-likelihood function of the data can be written as $L(\boldsymbol{Y}_n, \boldsymbol{\Omega}) = \sum_{t=1}^{n} \ln p(y_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\Omega})$, which, under normality, becomes

$$L(\boldsymbol{Y}_n, \boldsymbol{\Omega}) = \frac{-1}{2} \sum_{t=1}^{n} \Big[\ln(2\pi g_t^2) + \frac{(y_t - f_t)^2}{g_t^2} \Big].$$

Given prior distribution $p(\boldsymbol{\Omega})$, the log of the joint posterior distribution function for the model is

$$\ell(\Omega|\boldsymbol{Y}_n) \propto \ln[\, p(\boldsymbol{\Omega})] - \frac{1}{2} \sum_{t=1}^{n} \Big[\ln(2\pi g_t^2) + \frac{(y_t - f_t)^2}{g_t^2} \Big]. \tag{2}$$

The ability to evaluate this posterior function plays a key role in the proposed approach. For the general model in (1), this posterior function involves many parameters and might be difficult to handle. Some methods are available in the literature to overcome this difficulty, especially when special cases of model (1) are entertained. For example, the Kalman filter can be used to evaluate this posterior function recursively for linear Gaussian ARMA models with a flat prior (see Jones (1980)). The EM algorithm can be used if model (1) is in the form of a component model (see Shumway and Stoffer (1982)). More recently, the Gibbs sampler has been shown to be useful in obtaining the joint posterior distribution of $\boldsymbol{\Omega}$ for some time series models. For example, the Gibbs sampler with the Metropolis algorithm is found to be useful in modeling linear Gaussian ARMA models with conditionally conjugate priors. Here the Metropolis algorithm is used primarily to handle nonlinear parameters for which no closed-form formulas are available to simplify the Gibbs draw. In Carlin, Polson and Stoffer (1992), the Gibbs sampler in conjunction with scale mixtures of normal distributions was used to analyze nonlinear State-Space models. An advantage of the Gibbs sampler is that the joint posterior distribution of the model parameters in (2) can be obtained iteratively by using lower-dimensional conditional posterior distributions. As a special case, one may consider all 1-dimensional conditional posterior distributions in implementing the sampler. The 1-dimensional posterior distributions obtained from (2) are easy to evaluate. Another advantage of the Gibbs sampler is that only conditional prior specification is needed. Other Bayesian analyses of time series models using Markov Chain methods include Marriott et al. (1996) and Chib and Greenberg (1994) among many others.

In this paper, we also use the Gibbs sampler. However, we shall not use the Metropolis algorithm to handle nonlinear parameters. Instead we employ the griddy Gibbs approach of Tanner (1991) for those parameters that do not have closed-form formulas to facilitate the Gibbs draws. Advantages of the griddy Gibbs include simplicity and wide applicability. The computational burden of the griddy Gibbs, however, may be heavy. The griddy Gibbs used is as follows:

Griddy Gibbs.

Let ω_i be the ith element of $\boldsymbol{\Omega}$ and $[b_0, b_1]$ be the support of ω_i. In practice, the support is determined by properties of the entertained model. For example, if ω_i denotes the lag-1 coefficient of an AR(1) model, i.e. $y_t = \omega_i y_{t-1} + \epsilon_t$, then $[b_0, b_1] = [-1, 1]$ so that the process is not explosive. The conditional posterior distribution function of ω_i given the data, all the other parameters and prior distribution $p(\omega_i)$ is

$$p(\omega_i|\boldsymbol{Y}_n, \boldsymbol{\Omega}_{(i)}) \propto p(\omega_i|\boldsymbol{\Omega}_{(i)}) \prod_{t=1}^{n} N(f_t, g_t^2), \tag{3}$$

where $\boldsymbol{\Omega}_{(i)}$ denotes all the parameters in $\boldsymbol{\Omega}$ except ω_i and f_t and g_t are functions of ω_i. The griddy Gibbs draws a realization of ω_i by the following procedure:

- Select a grid of m points in the support $[b_0, b_1]$ or for a subset of $[b_0, b_1]$.
- For each grid point, evaluate the conditional posterior distribution function of ω_i in (3).
- Draw a random realization of ω_i from the selected grid based on the values of the conditional posterior distribution function.

From the procedure, it is clear that the actual value of the normalization constant of the conditional posterior function is not needed in implementing the griddy Gibbs. The prior distribution $p(\omega_i|\boldsymbol{\Omega}_{(i)})$ may assume many forms depending on the substantive information of the problem under study. It is clear, however, that a uniform prior simplifies the computation involved.

Example. As an illustration, we consider in detail the Gibbs sampler used for the following simple bilinear model $y_t = \phi_0 + \phi_1 y_{t-1} - \theta a_{t-1} + \beta y_{t-1} a_{t-1} + a_t$, $a_t = \sigma\epsilon_t$ where $\sigma > 0$ is the standard deviation of the innovation series a_t. The parameters of this model are $\boldsymbol{\Omega} = (\phi_0, \phi_1, \theta, \beta, y_0, a_0, \sigma)'$, where y_0 is the starting value of the series and a_0 is the starting innovation. The Gibbs samples of these parameters can be drawn as follows:

- The two AR coefficients ϕ_0 and ϕ_1 can be drawn easily because they are linear parameters and have a closed-form formula when a conjugate prior is used. Specifically, conditional on the other parameters, we can express the two AR parameters in a linear regression setup in a manner similar to that of an MA(1) model. Let $m_1 = \theta - \beta y_0$, $x_{11} = 1$ and $x_{21} = y_0$. For $t > 1$, define recursively $m_t = (\theta - \beta y_{t-1}) m_{t-1}$, $x_{1t} = 1 + (\theta - \beta y_{t-1}) x_{1,t-1}$ and $x_{2t} = y_{t-1} + (\theta - \beta y_{t-1}) x_{2,t-1}$. Furthermore, define $y_t^* = y_t - m_t$. Then, we have $y_t^* = \phi_0 x_{1t} + \phi_1 x_{2t} + a_t$, $t = 1, \ldots, n$. Therefore, ϕ_0 and ϕ_1 can be drawn jointly by using the usual result of Gibbs sampling for linear regression model with conjugate prior.

- The variance parameter σ^2 can also be drawn by using the usual technique, because conditional on other parameters σ^2 has an inverted chi-square distribution under the normality assumption and conjugate prior.
- The starting value y_0 can be drawn again by using results of linear regression analysis. Specifically, define $y_1^* = y_1 - \phi_0 + \theta a_0$, $x_1^* = \phi_1 + \beta a_0$ and $y_t^* = y_t - \phi_0 - \phi_1 y_{t-1} + (\theta - \beta y_{t-1}) y_{t-1}^*$ and $x_t^* = (\theta - \beta y_{t-1}) x_{t-1}^*$ for $t > 1$. Then, we have $y_t^* = x_t^* y_0 + a_t,\ t = 1, \ldots, n$, and the result of the Gibbs sampler for simple linear regression applies.
- Similarly, the starting innovation a_0 can be drawn by using the result of linear regression analysis. Define $y_1^* = y_1 - \phi_0 - \phi_1 y_0$, $x_1^* = -\theta + \beta y_0$, $y_t^* = y_t - \phi_0 - \phi_1 y_{t-1} + (\theta - \beta y_{t-1}) y_{t-1}^*$ and $x_t^* = (\theta - \beta y_{t-1}) x_{t-1}^*$ for $t > 1$. Then, we obtain $y_t^* = x_t^* a_0 + a_t,\ t = 1, \ldots, n$, which is a simple linear regression.
- Finally, the MA coefficient θ and the bilinear parameter β are nonlinear, and there exist no closed-form formulas to simplify the Gibbs draw. One possible approach to overcome this difficulty is to use the Metropolis algorithm. In this paper, we use the griddy Gibbs approach. As mentioned before, for these two parameters, the individual conditional posterior distribution functions can be evaluated easily over a grid of finite points. For the MA coefficient θ, the support is $[-1, 1]$ whereas that of the bilinear parameter β must satisfy the condition $\phi_1^2 + \sigma^2 \beta^2 < 1$. (See Liu (1989) for the stationarity condition of the bilinear model.)

Note that in theory all parameters can be drawn by using the griddy Gibbs. However, it is desirable to use closed-form formulas whenever available and to draw several parameters jointly whenever possible. Drawing one parameter at a time using the griddy Gibbs could result in slow convergence of the sampler.

In the above illustration, all techniques used are not limited to the bilinear model. On the contrary, the proposed estimation procedure is widely applicable in linear and nonlinear time series analysis. Only the closed-form formulas and the likelihood function need to be changed when other models are entertained.

A potential weakness in using the griddy Gibbs is the specification of parameter support. For simple models, one can use the theoretical properties of the model such as stationarity, invertibility or existence of some moments to select the supports. However, for high dimensional models, the interdependence of the parameters may complicate the specification. In our implementation of the griddy Gibbs, we use an iterative method. We start with a relatively wide interval for a given parameter and refine the interval after some Gibbs iterations. In practice, this means one needs to run the Gibbs sampler several times in order to obtain estimates of a model. Given the advance in computing facilities and the gains in understanding the series over the iterations, we believe that this is not a serious drawback for the proposed estimation method. Furthermore, when

the number of parameters is large, one can start with a sparse grid in the initial Gibbs iterations to reduce the computation in refining the specification of parameter supports.

4. Model Discrimination

In this section we consider the problem of model selection in nonlinear time series analysis, especially when the competing models are not nested. Such a model-selection problem is important because many classes of nonlinear models have been proposed in the literature and there exists no simple method to effectively discriminate one class of models from another. For example, both the TAR and bilinear models have been used to analyze annual sunspot data with proponents claiming better fit for their model. (See Tong (1990) and Gabr and Subba Rao (1981).) Another example is that many ARCH-type of models have been used to describe and predict the volatility of the monthly S&P 500 excess returns, and there is no agreement on which model is most appropriate.

Our approach to model discrimination is to let individual observations make their own choice of model. Consider the case of two competing non-nested nonlinear models. We use these competing models to define the functions $f(\cdot)$ and $g(\cdot)$ of model (1) and introduce a simple switching scheme that allows each individual observation to select its own model. Thus, under the proposed approach, the two competing models become submodels of a mixed model, and each individual observation can select its own submodel. In real applications, the dynamic structure of a time series cannot change abruptly over time. A structural change tends to occur gradually over a period of time. Therefore, it is reasonable to assume that the model selection of individual observations evolves over time in a smooth fashion. This consideration leads us to employ a simple switching scheme to govern the model selection. The selection results of individual observations provide information about which submodel is more appropriate for the data. This information can then be used to make model selection.

The idea behind the proposed mixed model is simple. We believe that the issue of model discrimination exists only when the two competing non-nested nonlinear models fit the data well; otherwise, the selection is clear. Consequently, a better way to discriminate between models is to let each individual observation select its own model. Moreover, it is conceivable that certain portions of the data fit one model nicely whereas the remaining data fit the other model better. In this situation, the mixed model considered appears to be more appropriate.

One can also treat the proposed mixture-model approach as a generalization of the odds-ratio commonly used in Bayesian inference. In computing an odds-ratio, we assume that *all* of the data points belong to the candidate model. On the other hand, under the proposed mixed model, observations can belong to

different models. Thus, the proposed method provides another level of flexibility over the odds-ratio.

The switching scheme for model discrimination has been used in McCulloch and Tsay (1994) to test for "trend-stationarity" versus "difference-stationarity" of a linear time series and in George, McCulloch and Tsay (1996) to distinguish between fixed-coefficient versus random-coefficient autoregressive models. However, these two papers use Markov switching and only consider linear models for which closed-form formulas are available. The current paper is much more general as it can handle a wide range of linear and non-linear models. In addition, this paper improves the procedure by treating starting values and innovations as parameters . This improvement could be significant in applications because it relaxes the assumption that the starting values and innovations are either fixed or equal to their expectation. For non-stationary series to which most real-world time series belong, the unconditional expection of a series might not exist.

The probabilistic mechanism of the proposed mixed model can give rise to a large number of possible submodel configurations; for a given time series of length n, the possible number of submodel configurations is 2^n. In applications, these configurations might require intensive computation in model estimation. However, as illustrated in George and McCulloch (1993), McCulloch and Tsay (1994) and George, McCulloch and Tsay (1996), this computational difficulty can be overcome by using Gibbs sampling.

4.1. A mixed model with switching

The proposed framework for discriminating between two competing models is the two-state switching model:

$$y_t = \begin{cases} f_{1,t} + a_{1t}, & a_{1t} = g_{1,t}\epsilon_t \quad \text{if } s_t = 1, \\ f_{2,t} + a_{2t}, & a_{2t} = g_{2,t}\epsilon_t \quad \text{if } s_t = 2, \end{cases} \tag{4}$$

where $f_{i,t}$, $g_{i,t}$ and a_{it} are defined as in (1) and $\{s_t\}$ is a sequence of states. The state switching is governed by

$$P(s_t = i | s_{t-1} = s_{t+1} = i) = \eta, \quad P(s_t = i | s_{t-1} \neq s_{t+1}) = 0.5. \tag{5}$$

Thus the switching depends on the two nearest neighbors of the observation in time. Each observation has a conditional probability η to stay with the same model as its neighbors. When the two adjacent neighbors are in different states, the probability of model switch is neutral at 0.5. Because structural changes tend to occur gradually over a period of time, a large η seems to be more realistic in application. If $\eta = 1$ then all observations come from one of the two competing models. In this paper, we use η close to 1 and consider values of η in the interval

[.95, .9999] to study the sensitivity of model selection with respect to the choice of η. Alternatively, one could put a hyper-prior on η that has most of the probability mass on values close to 1. Including the large set of parameters s_t in model (4) makes it very flexible. In application strong prior information (large η) is needed to get reasonable results.

One way to appreciate the implication of η is to consider the independence case in which model change is independent over time. In this case, there are 8 possible model configurations for every three consecutive observations and the probability that all three observations belong to the same model is only $\eta/4$, which is less than 0.25. Thus, the chance of model change is substantial when η is not large. Of course, the independence assumption is an extreme case and is often unrealistic in application. Our discussion is only meant to justify the use of a large value for η.

There are many ways to describe the transition of model selection from one observation to another, ranging from independent Bernoulli trials to complicated dynamic mechanism. Our choice of (5) is based on several considerations. First, the transition is very flexible; it covers a wide range of possibilities by varying η. For example, $\eta = 0.5$ corresponds to independent Bernoulli trials with probability 0.5 and $\eta = 1$ implies that change can only occur when two neighbors belong to different models. Second, it is easy to use because the user only needs to specify a single parameter. In the traditional two-state Markov switching model, one needs to specify two parameters for the probability transition matrix. Third, the equation is intuitively appealing. It easily reflects the common sense of smooth model change. Fourth, the scheme can be extended to involve other neighboring systems, e.g. two observations prior and after the observation. Such a specification would provide an alternative way to specify strong prior information that nearby observations are likely to come from the same model.

4.2. Implementation

Model selection is based on the posterior distribution of the parameters s_t in model (4). This posterior is computed in the obvious way by using Gibbs sampling and drawing the s_t's given the parameters of both models and then drawing the parameters of the individual models given the s_t values in a manner similar to that outlined in Section 3.

To use our method for model discrimination, we propose the following procedure:

1. For each submodel we specify prior distributions for the model parameters and then use all the data and the estimation method of Section 3 to obtain estimates (typically posterior means).

2. Choose a value for η and perform a Gibbs estimation of the mixed model in (4). Initial values for the submodel parameters are obtained from step 1. Step 1 also provides an initial choice of the support and grid for each parameter drawn using the griddy method.
3. Check the convergence of Gibbs sampler. Refine and iterate Gibbs sampler if necessary.
4. Use the posterior distribution of model selection of individual observations to make inference.

Once the posterior distribution of individual selection is available, one can make inference of model selection based on the objective of the analysis. For example, if the objective is forecasting, one may pay more attention to the model selection of observations close to the forecast origin. If the objective is the dynamic structure of the data, then the posterior mean or median can be used for overall selection. It is conceivable that the data might not have sufficient information to distinguish one competing model from another. In this case, one might search for more data or for ways to further improve the model. It would be unwise to assume that a statistical method can always distinguish two competing models based on a finite sample of observations.

Finally, it is important in practice to study the sensitivity of model selection with respect to prior specification such as η and to check the convergence of the Gibbs sampler. By varying priors and the number of iterations and starting values of the Gibbs sampler, one can learn the stability of model selection.

5. Examples

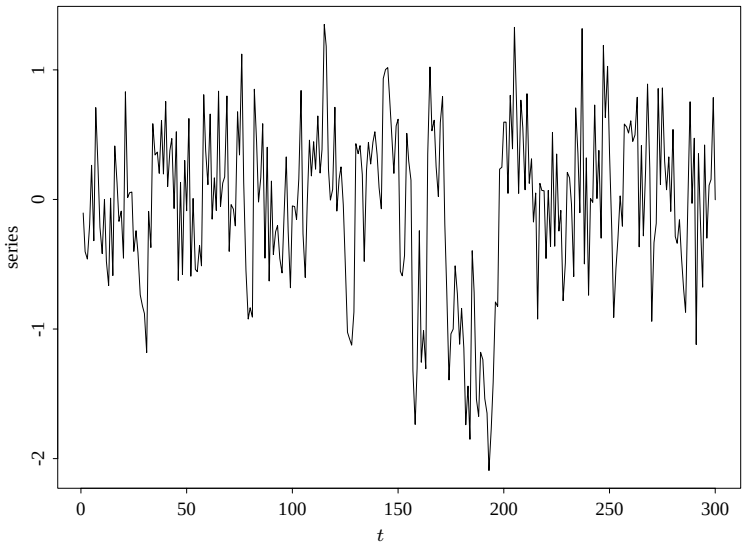
We illustrate the proposed unified approach to estimating and modeling time series by some simulated and real examples. For the simulated examples, we consider AR versus MA models and TAR versus bilinear models. We analyse two real data sets. We compare TAR and bilinear models for the annual sunspot numbers and ARCH(2) and GARCH(1,1) models for monthly excess returns of the S&P500 stock market portfolio.

Example 1. Figure 1(a) shows a time plot of 300 observations generated from the model

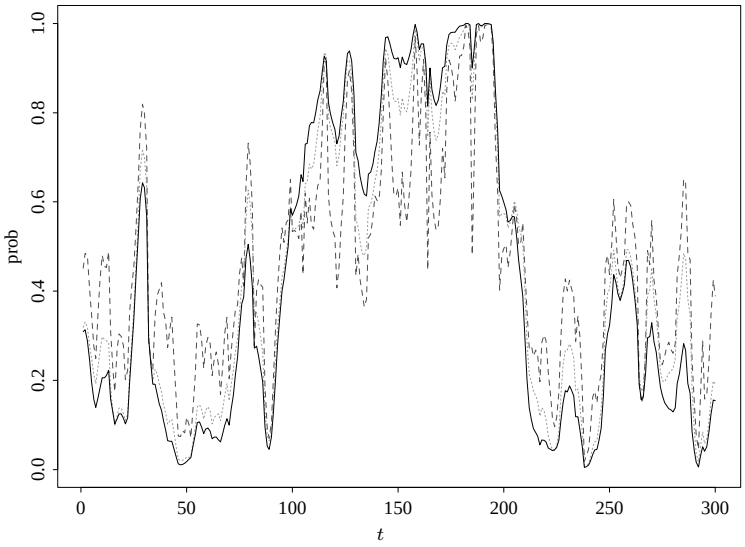
$$y_t = \begin{cases} .8y_{t-1} + a_t, & \text{if } t = 101, \ldots, 200, \\ a_t + .3a_{t-1} + .4a_{t-2}, & \text{if } t = 1, \ldots, 100; 201, \ldots, 300, \end{cases}$$

where $a_t = 0.5\epsilon_t$, $a_0 = a_{-1} = 0$ and $y_0 = 0$. This is a mixed model with two change points at $t = 101$ and $t = 201$. The model change at $t = 201$ can be seen in Figure 1(a), but that at $t = 101$ is not obvious. Our goal here is to illustrate the performance of the proposed approach in estimation and model selection. For

the AR model, the parameter vector is $\boldsymbol{\Omega}_1 = (\phi_1, y_0, \sigma_1)'$, where ϕ_1 is the lag-1 AR coefficient and σ_1 is the standard deviation of the AR innovation. For the MA model, the parameter vector is $\boldsymbol{\Omega}_2 = (\theta_1, \theta_2, a_0, a_{-1}, \sigma_2)'$, where θ_i are MA coefficients, a_i are starting innovations and σ_2 denotes the standard deviation of the MA innovation.



(a) simulated series



(b) probability of MA model

Figure 1. Time plot and posterior probabilities for Example 1.

Following the proposed procedure in Section 4.2, we began with Gibbs samples for each model, assuming that all of the data belong to that model, to obtain initial parameter estimates and the initial parameter supports for the griddy Gibbs. The prior for any parameter drawn by griddy Gibbs is the uniform distribution over its interval support. The initial Gibbs samples used 300 iterations. Using results of the initial Gibbs samples and a given η for conditional switching probability, we ran 2500 Gibbs iterations to obtain posterior distribution of individual model selection. The estimated posterior probabilities are based only on the last 2000 iterations. This step of ignoring the first 500 Gibbs iterations was taken to reduce the effect of initial parameter specification. Figure 1(b) shows the posterior mean of selecting the MA model for each individual observation for $\eta = 0.95, 0.99, 0.995$, respectively. The solid line is for $\eta = 0.995$ and the dotted line for $\eta = 0.99$. The effect of η on model selection is seen from the three posterior probabilities. As expected, $\eta \geq 0.99$ works better and is preferred. For this example, it is seen that the proposed model-selection method works reasonably well. It points out clearly the two change points and is able to identify the generating model. For estimation, the posterior distributions of the parameters are well behaved and are centered roughly around the true values.

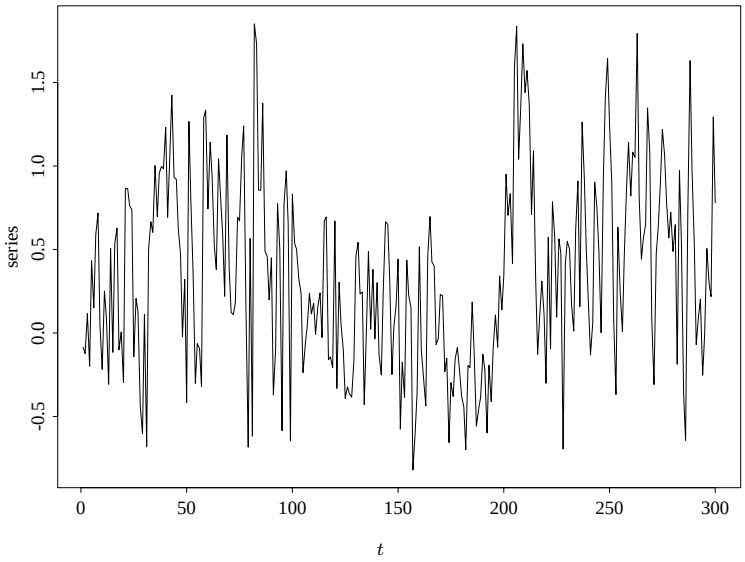
Example 2. In this example, we generated 300 observations from the mixed model

$$y_t = \begin{cases} \begin{cases} .8y_{t-1} + a_{1t} & \text{if } y_{t-1} \geq 0 \\ -.8y_{t-1} + a_{1t} & \text{if } y_{t-1} < 0 \end{cases} & \text{if } t = 1, \ldots, 100; 201, \ldots, 300 \\ .5y_{t-1} + .2y_{t-1}a_{2,t-1} + a_{2t} & \text{if } t = 101, \ldots, 200 \end{cases}$$

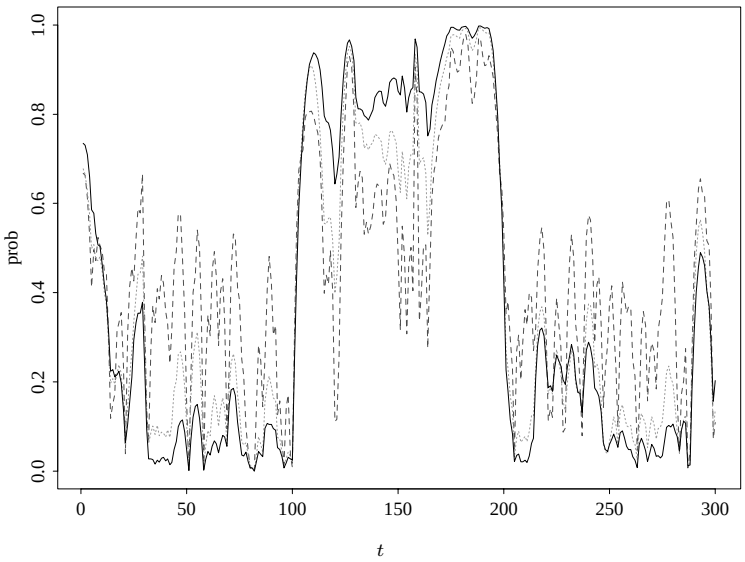
where $y_0 = a_0 = 0$, $a_{1t} = 0.5\epsilon_t$ and $a_{2t} = 0.3\epsilon_t$. This is a mixture of TAR and bilinear models with two change points at $t = 101$ and $t = 201$. The TAR model has two regimes separated by the threshold variable y_{t-1} at threshold $r = 0$. In each regime, the model is AR(1). The bilinear model used contains a single bilinear term $.2y_{t-1}a_{t-1}$ and is referred to as a "diagonal" bilinear model. A special feature of such a bilinear model is that the mean of the series is nonzero, even though there is no constant term in the model. Properties of diagonal bilinear models are more complicated than those of non-diagonal bilinear models. (See Guegun (1994).)

The data of this example are shown in Figure 2(a). Even a careful reading of the plot cannot reveal easily the two change points. In our analysis, we assume that the threshold variable y_{t-1} is known, but the threshold r is unknown. Thus, the parameter vector for the TAR submodel is $\boldsymbol{\Omega}_1 = (r, \phi_1^{(1)}, \phi_1^{(2)}, \sigma_1, y_0)'$ where r denotes the threshold , $\phi_1^{(i)}$ is the AR(1) coefficient of the ith regime, and σ_1 is the innovational standard deviation. For the bilinear submodel, the parameter

vector is $\boldsymbol{\Omega}_2 = (\phi, \beta, \sigma_2, y_0, a_0)'$, where ϕ and β are the AR and bilinear coefficient, respectively, σ_2 denotes the standard deviation of innovations, and y_0 and a_0 denote the starting value and innovation, respectively. In sum, there are 10 parameters in the mixed model used for model selection.

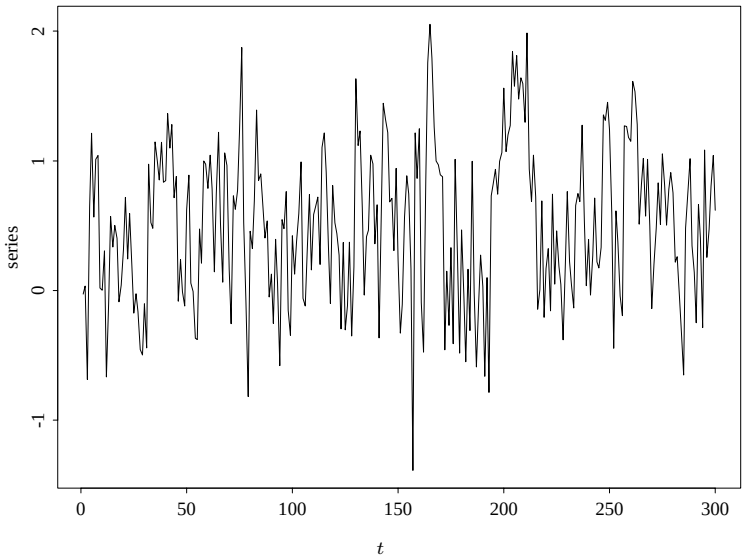


(a) simulated series

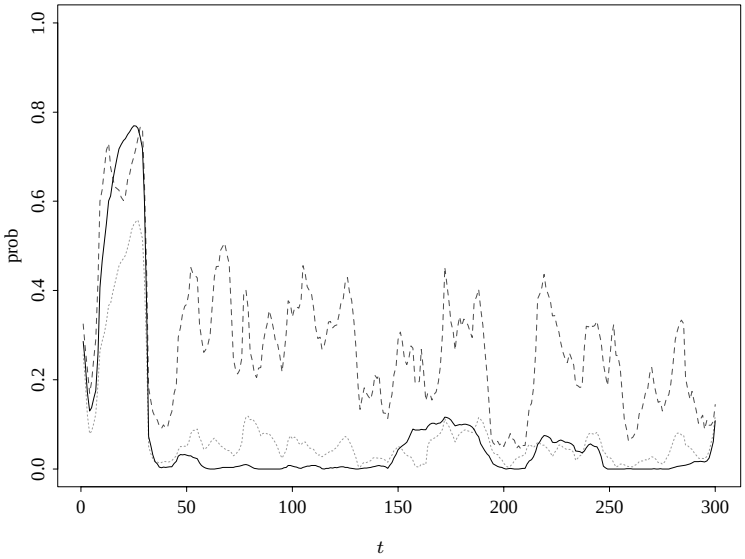


(b) probability of bilinear model

Figure 2. Time plot and posterior probabilities for Example 2.



(a) simulated series



(b) probability of bilinear model

Figure 3. Time plot and posterior probabilities for Example 3.

Following the proposed procedure of Section 4.2 and using essentially the same Gibbs steps and numbers of iterations as those of Example 1, we obtain the posterior probability of selecting the bilinear model for each observation. These

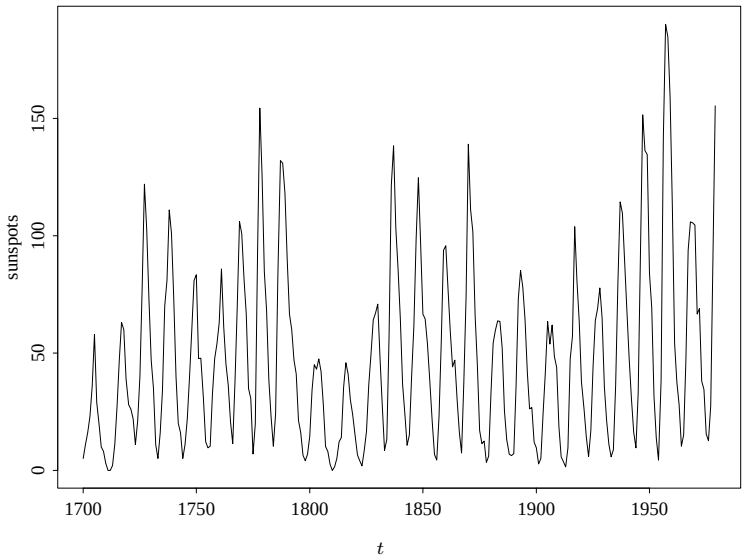
probabilities are shown in Figure 2(b) for $\eta = 0.95, 0.99, 0.999$. It is seen that the proposed procedure works well for the data, especially for $\eta = 0.999$. The two change points and the generating model are clearly identified. The result for $\eta = 0.95$ given by the dashed line in Figure 2(b) shows some model uncertainty at some patches of observations. This is reasonable because the prior probability for model change in this case is substantial.

Example 3. In this example, we generated 300 observations all from the TAR submodel of Example 2. However, we assume that the bilinear submodel is another competing model and apply the proposed procedure to discriminate these two models. Figure 3(a) shows the data whereas Figure 3(b) gives the posterior probabilities of selecting the bilinear model by the individual observations. These probabilities were obtained by using the same starting values, the same Gibbs steps and iterations as those of Example 2, except that the probability of conditional model switching is set at $\eta = 0.99$, 0.999, and 0.9999, respectively. From the probability plot, it is seen that the proposed procedure indeed selects the generating model for the data, especially when η is close to 1. The case of $\eta = 0.99$ shows some model uncertainty, even though only some isolated points have posterior probability greater than 0.5 for the bilinear model. This example thus shows that the prior specification of η should be close to 1 in applications, say $\eta \geq 0.99$.

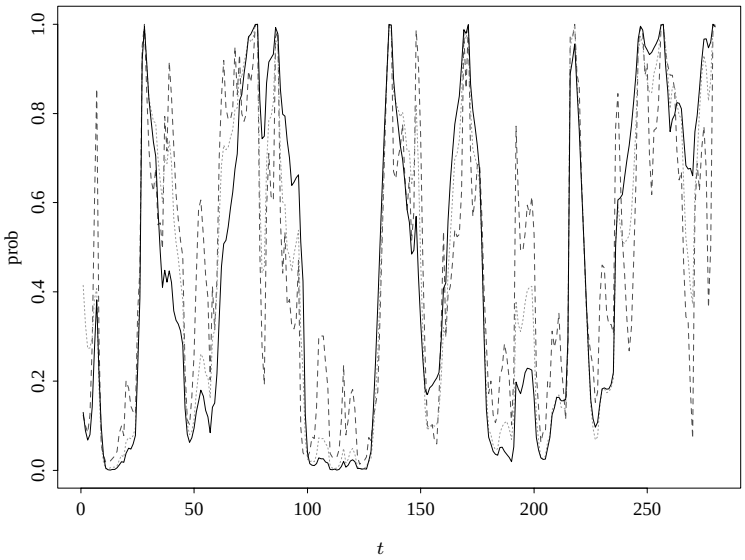
Example 4. In this example, we consider the annual Wolf sunspot number from 1700 to 1979 for 280 observations. The data shown in Figure 4(a) are listed in Tong (1990) and have been widely used in nonlinear time series analysis. It is generally believed that this series is nonlinear, but there is no agreement on which nonlinear model is most appropriate for the data. When the subsample from 1700 to 1921 was used, Gabr and Subba Rao (1981) identified a bilinear model for the series whereas Tong (1990) specified a two-regime TAR model. Li (1993) applied a test statistic, which uses the idea of separate families of hypotheses of Cox (1962), to the subsample and concluded that the bilinear model of Gabr and Subba Rao is more appropriate. However, from a theoretical view point, bilinear models do not possess the asymmetric feature between rise and fall of the cyclical pattern observed in the sunspot number. On the other hand, the TAR is capable of producing an asymmetric cycle, but it has larger residual variance in the subsample. The issue of model selection remains.

Our analysis here is to apply the proposed model-discrimination procedure to the full sample, assuming that the bilinear model of Gabr and Subba Rao (1981) and the TAR of Tong (1990) as two competing models. The bilinear model considered assumes the form:

$$\begin{aligned} y_t = {} & \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_9 y_{t-9} + \beta_{21} y_{t-2} a_{t-1} + \beta_{81} y_{t-8} a_{t-1} + \beta_{13} y_{t-1} a_{t-3} \\ & + \beta_{43} y_{t-4} a_{t-3} + \beta_{16} y_{t-1} a_{t-6} + \beta_{24} y_{t-2} a_{t-4} + \beta_{32} y_{t-3} a_{t-2} + a_t, \end{aligned} \tag{6}$$



(a) sunspots



(b) probability of TAR model

Figure 4. Time plot and posterior probabilities for annual sunspot numbers.

where $a_t = \sigma\epsilon_t$. Besides the 12 parameters shown in equation (6), this bilinear model also needs 9 starting values $y_0, y_{-1}, \ldots, y_{-8}$ and 6 starting innovations $a_0, \ldots, a_{-5}$. In total, estimation of this bilinear model considers 27 parameters

some of which are highly nonlinear. As shown by the simpler bilinear example of Section 3, we can estimate this bilinear model via the proposed Gibbs sampler. The Gibbs draws of the nonlinear parameters can be done by the griddy Gibbs.

The threshold model built by Tong (1990) is

$$y_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{3} \phi_i^{(1)} y_{t-i} + a_t^{(1)}, & \text{if } y_{t-3} \le r, \\ \phi_0^{(2)} + \sum_{i=1}^{11} \phi_i^{(2)} y_{t-i} + a_t^{(2)}, & \text{if } y_{t-3} > r, \end{cases} \tag{7}$$

where $a_{it} = \sigma_i \epsilon_t$. Including innovational standard deviation, this TAR model contains 5 parameters in Regime 1 and 13 parameters in Regime 2. Counting the threshold r and 11 starting values $y_0, y_{-1}, \ldots, y_{-10}$, we are effectively estimating 30 parameters for the TAR model. Except for the threshold r, all of the parameters have closed-form formulas and can be drawn easily. Conditioned on other parameters, the threshold r becomes a change point of the data. Gibbs draws of r, therefore, can be done by either the griddy Gibbs or the method in Carlin, Gelfand and Smith (1992).

Again, we follow the proposed procedure in Section 4.2 to carry out the model selection. Due to the large number of parameters involved, we used 3500 Gibbs iterations for this example, but discarded results of the first 500 iterations in computing the posterior probabilities. Figure 4(b) shows the posterior probabilities of selecting the TAR model by the individual observations, where the solid, dashed, and dotted lines are for $\eta = 0.99, 0.999$ and 0.9999, respectively. In our analysis, we carried out many Gibbs samples and found that the posterior probability plot is stable. From the plots, it is seen that the data do not strongly favor a single model. For certain periods, the TAR was preferred. But for other periods, the bilinear model was selected. It seems that the data are not sufficiently informative to discriminate between these two competing models. However, the TAR model appears to be the choice of model by the most recent observations. This is in good agreement with the results of forecasting comparison in Tong (1990), Sec. 7.3 who showed that the TAR model produced better out-of-sample forecasts of the sunspot numbers for the latter part of the data.

The fact that the data were not very informative in choosing a single model is understandable. First, the two competing models entertained contain many parameters, making them rather flexible and capable of providing good fit in finite samples. In this circumstance, one might need a large number of observations to distinguish one model from the other. Second, there exists the possibility that neither of the two competing models are appropriate for the data. This is evident in the posterior probability plot of Figure 4(b) where the TAR model was preferred when the sunspot number was high and the bilinear model was chosen when the sunspot number was low. In addition, our residual analysis shows that

the normalized residuals of the mixed model has lag-1 serial correlation, even though the correlation is relatively weak.

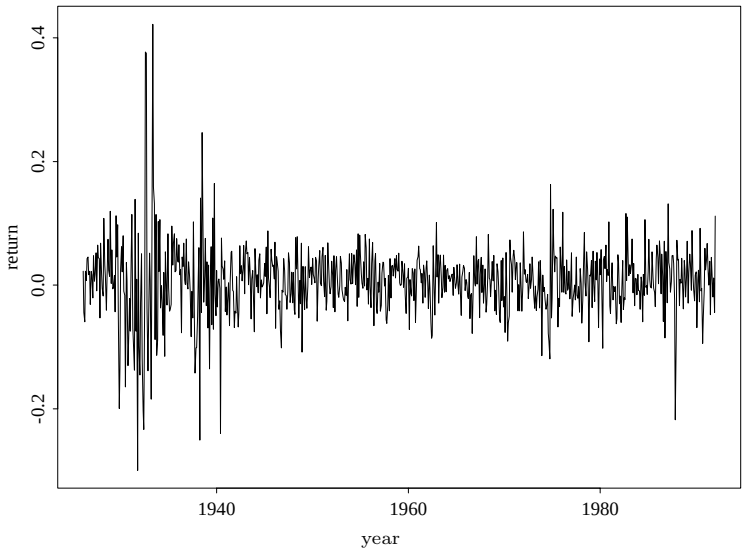
In summary, the proposed procedure does not pinpoint a single model for the annual sunspot number from 1700 to 1979. However, it produces results that are reasonable and in agreement with those available in the literature. It would be unwise to expect that a model-selection method can always select a single model based on a finite number of observations. One must take into consideration the possibility that there exists no true model for a real-world time series. When the data are not sufficiently informative, a good model-selection procedure should be able to reveal it. In this sense, the proposed model-discrimination method appears to be reasonable.

Example 5. Figure 5(a) is a time plot of monthly excess returns of the S&P500 portfolio from January 1926 to December 1991 giving 792 observations. This series has been widely analyzed in volatility studies, but there is little agreement on what is the most appropriate model for the data. Our goal here is to compare between ARCH(2) and GARCH(1,1) models for the data. The model considered is

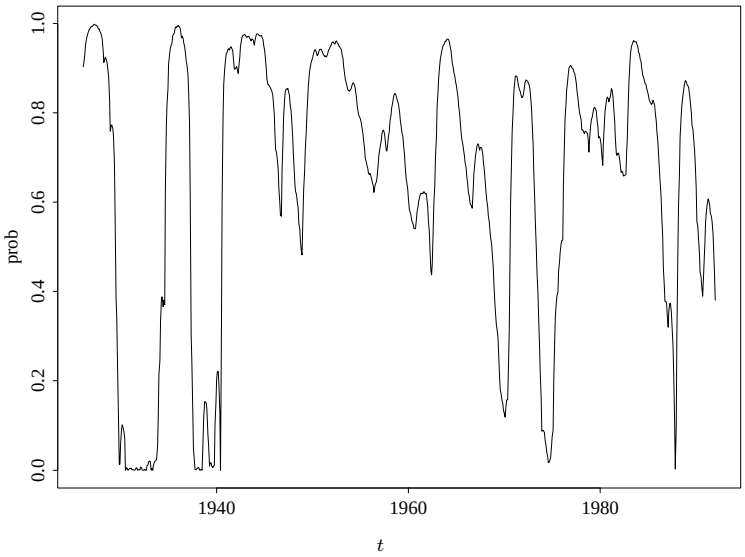
$$y_t = \begin{cases} \theta_0 + a_{1t}, & a_{1t} = g_{1,t}\epsilon_t, \quad g^2_{1,t} = \gamma_0 + \gamma_1 a^2_{1,t-1} + \gamma_2 a^2_{1,t-2} & \text{if } s_t = 1 \\ \beta_0 + a_{2t}, & a_{2t} = g_{2,t}\epsilon_t, \quad g^2_{2,t} = \alpha_0 + \alpha_1 g^2_{2,t-1} + \beta_1 a^2_{2,t-1} & \text{if } s_t = 2. \end{cases}$$

The state switching is governed by equation (5) with $\eta = 0.999$. Figure 5(b) plots the posterior probability of the ARCH(2) model based on the Gibbs sampler with 500 initial iterations and 4000 general iterations. The mean of the posterior probabilities is 0.65 so that the overall fit is slightly in favor of the ARCH(2) model. On the other hand, by comparing Figures 5(a) and 5(b), the GARCH(1,1) model was selected by most observations that appear to be volatile. Thus, our result indicates that the evidence of GARCH(1,1) model reported in the literature is largely due to the few visibly volatile periods of the US economy. While such a conclusion is understandable, the proposed analysis does highlight the influential periods for using GARCH(1,1) model. This shows that the proposed model discrimination procedure can be used to monitor the evolution of the time series under study.

In the estimation, we used various constraints to ensure that the two sub-models have proper unconditional variances. For instance, we require $\alpha_0 > 0$, $0 \leq \alpha_1 + \beta_1 < 1$, $\alpha_1 \geq 0$, and $\beta_1 \geq 0$ so that the GARCH(1,1) model is not integrated. Such constraints are easy to implement under the proposed unified approach.



(a) Monthly SP500 excess returns



(b) probability of ARCH(2) model

Figure 5. Time plot and posterior probabilities for monthly S&P500 returns.

Finally, all computations in this paper were done by a program written in C^{++}. This program can be used for the general model in (1) provided that the $f(\cdot)$ and $g(\cdot)$ functions are given.

Acknowledgement

This research is supported in part by the National Science Foundation, the Graduate School of Business, University of Chicago and the Ministry of Education, Republic of China. The original version of this work was done when C. Chen was visiting the Graduate School of Business, University of Chicago.

References

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive moving average process. *Biometrika* **66**, 59-65.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31**, 307-327.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd Edition. Prentice-Hall, Englewood Cliffs, New Jersey.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, 2nd Edition. Springer-Verlag, New York.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of change point problems. *Appl. Statist.* **41**, 389-405.

Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992). A monte carlo approach to nonnormal and nonlinear state-space modeling. *J. Amer. Statist. Assoc.* **87**, 493-500.

Chib, S. and Greenberg, E. (1994). Bayes inference in regression models with ARMA(p,q) errors. *J. Econometrics* **64**, 183-206.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Roy. Statist. Soc. Ser. B* **24**, 406-424.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987-1007.

Gabr, M. M. and Subba Rao, T. (1981). The estimation and prediction of subset bilinear time series models with applications. *J. Time Ser. Anal.* **2**, 155-171.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881-889.

George, E. I., McCulloch, R. E. and Tsay, R. S. (1996). Two approaches to Bayesian model selection with applications. In *Bayesian Analysis in Statistics and Econometrics* (Edited by D. A. Berry, K. M. Chaloner and J. K. Geweke). John Wiley, New York.

Guegan, D. (1994). *Series Chronologiques Nonlinear a Temps Discret*. Economica, Paris.

Granger, C. W. J. and Andersen, A. P. (1978). *Introduction to Bilinear Time Series Models*. Vandenhoeck and Ruprecht, Gőttingen.

Hillmer, S. C. and Tiao, G. C. (1979). Likelihood function of stationary multiple autoregressive moving average models. *J. Amer. Statist. Assoc.* **74**, 652-660.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**, 389-395.

Li, W. K. (1993). A simple one degree of freedom test for non-linear time series model discrimination. *Statist. Sinica* **3**, 245-254.

Liu, J. (1989). A simple condition for the existence of some stationary bilinear time series. *J. Time Ser. Anal.* **10**, 33-39.

Marriott, J., Ravishanker, N., Gelfand, A. and Pai, J. (1996). Bayesian analysis of ARMA processes: complete sampling-based inference under exact likelihoods. In *Bayesian Analysis in Statistics and Econometrics* (Edited by D. A. Berry, K. M. Chaloner and J. K. Geweke). John Wiley, New York.

McCulloch, R. E. and Tsay, R. S. (1994). Bayesian inference of trend- and difference-stationarity. *Econom. Theory* **10**, 596-608.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press. New York.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.* **3**, 253-264.

Subba Rao, T. (1981). On the theory of bilinear time series models. *J. Roy. Statist. Soc. Ser. B* **43**, 244-255.

Tanner, M. A. (1991). *Tools for Statistical Inference*. Springer-Verlag, New York.

Tong, H. (1978). On a threshold model. In *Pattern Recognition and Signal Processing* (Edited by C. H. Chen). Sijhoff Noordhoff, Amsterdam.

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

Tsay, R. S. (1987). Conditional heteroscedastic time series models. *J. Amer. Statist. Assoc.* **82**, 590-604.

100 Wenhua Rd., Hsitun District, Taichung 40724, Taiwan.

Graduate School of Business, University of Chicago, 1101 E. 58th St., Chicago, IL 60637, U.S.A.

(Received September 1992; accepted October 1996)