Nonlinear Time Series Modelling: An Introduction[*]

Simon M. Potter
Federal Reserve Bank of New York
33 Liberty St.
New York, NY, 10045
email: simon.potter@ny.frb.org

August 1999

Abstract

Recent developments in nonlinear time series modelling are reviewed. Three main types of nonlinear model are discussed: Markov Switching, Threshold Autoregression and Smooth Transition Autoregression. Classical and Bayesian estimation techniques are described for each model. Parametric tests for nonlinearity are reviewed with examples from the three types of model. Finally forecasting and impulse response analysis is developed.

Keywords: Markov Switching, Threshold Autoregression, Smooth Transition Autoregression.

1 INTRODUCTION

It is now ten years since Jim Hamilton's seminal paper on nonlinear modelling of U.S. output was published. This ten years has witnessed an explosion of interest amongst econometricians in the testing, estimation, specification and properties of nonlinear models. The purpose of this paper is to give a non-technical survey of the main developments and some observations on the difficulties of successful nonlinear modelling in macroeconomics.[1] It

[*] I would like to thank Gary Koop for numerous helpful discussions. The views expressed in this paper are those of the author and do not necessarily reflect the views of the Federal Reserve Bank of New York or the Federal Reserve System.

[1] There is also a vast literature in finance (see De Lima's accompanying article). While many of the developments and problems are similar, the higher frequency of observation of financial time series has allowed a much greater emphasis on flexible non-parametric methods. An excellent example of such methods is in Gallant, Rossi and Tauchen (1993).

is useful to begin by giving some motivation for the need for nonlinear modelling.

The 1970s and 1980s saw economists adopt many of the time series techniques introduced by Box and Jenkins. The basis for such modelling approaches was the Wold Representation: any covariance stationary time series can be expressed as moving average function of present and past innovations:

$$Y_t = \sum_{i=0}^{\infty} \theta_i U_{t-i}, \text{with } \sum \theta_i^2 < \infty,\ \theta_0 = 1,$$

where

$$E[U_t U_{t-i}] = 0 \text{ for all } i \neq 0 \text{ and } E[U_t^2] = \sigma_u^2$$

This infinite moving average can nearly always be well approximated by low order autoregressive processes perhaps with some moving average components. Further, the dynamics of the time series could be 'read off' from the Wold Representation since θ_i represents the impulse response function at horizon i.

It might appear at first that there is no need for nonlinear modelling given the Wold Representation. However, as well illustrated in the appropriately titled "Forecasting White Noise," by Clive Granger, lack of autocorrelation in a time series does not imply that the time series cannot be predicted. Indeed some perfectly predictable time series have zero autocorrelations at all lags.[2]

Further, while the Wold Representation gives the impulse response function directly it imposes some strong restrictions on it. First, the impulse response function does not depend on the recent history of the time series. Thus, for example, the response to a positive innovation of 1% is the same whether last period's growth rate was 8% or -8%. Second, the response to innovations is restricted to be homogeneous of degree 1. That is, once the response to a shock of size 1 has been found all other shocks are given by simple scalings of this response.

Successful nonlinear time series modelling would improve forecasts and produce a richer notion of business cycle dynamics than linear time series

[2] The classic example is Brock and Chamberlain's 1984 working paper which like Granger's paper has a title that gives the result. In the late 1980s nonlinear modeling was strongly associated with the study of chaotic systems. Such systems are less amenable to statistical techniques than the nonlinear time series models considered here.

models allow. For this to happen two conditions are necessary. First, economic time series must contain nonlinearities. Second, we need reliable statistical methods to summarize and understand these nonlinearities suitable for time series of the typical macroeconomic length. Unfortunately the second condition is needed to evaluate the veracity of the first condition and as we shall see it is not clear that we have yet found reliable statistical methods.

The organization of the paper is as follows: I start by describing three types of models most widely used in the economics literature and Classical and Bayesian estimation techniques in simple cases; the testing problem is then discussed with respect to these three models; finally simulation of the conditional expectations is described and its use in the construction of forecasts and impulse response functions.

2 THREE MODELS

In this section I discuss the three types of models that have most commonly be used in nonlinear modelling particularly for aggregate output measures and unemployment. I will use a common notation across all models. Y_t will be a univariate covariance stationary time series, $Y^t = (Y_1, Y_2, \ldots, Y_t)$ will be the history of the time series up to time t. V_t will be a sequence of independent and identically distributed random variables with unit variance. When likelihood based methods are discussed one can assume that V_t has a standard normal distribution. The Greek letters α, ϕ, σ will respectively refer to the intercepts, autoregressive coefficients and the scaling of the time series innovation. $\phi(L)$ is a polynomial in the lag operator of the form:

$$\phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p.$$

Below we will make use of the fact that if $V_t \sim N(\mu, 1)$ and the prior belief on μ is flat over the line then the posterior belief about μ after observing a sample of size T is

$$\mu \sim N(\frac{1}{T}\sum_{t=1}^{T} v_t, \frac{1}{T}).$$

2.1 Markov Switching

It is best to begin with Hamilton's model from Econometrica 1989. His original motivation was to model long swings in the growth rate of output but instead he found evidence for discrete switches in the growth rate at business cycle frequencies. Output growth was modelled as the sum of a discrete Markov chain and a Gaussian autoregression:

$$Y_t = Z_t + X_t,$$

where

$$Z_t = \alpha_0 + \alpha_1 S_t, S_t = 0 \text{ or } 1$$

and

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \phi_4 X_{t-4} + \sigma V_t,$$

and $P[S_t = 1 | S_{t-1} = 1] = p, P[S_t = 0 | S_{t-1} = 0] = q, V_t \sim N(0, 1)$.

The major estimation difficulty with the model is the lack of separate observability of Z_t and X_t. A simple variation on the model is:

$$Y_t = Z_t + \phi(L) Y_{t-1} + \sigma V_t,$$

where only S_t is now unobservable. Note that the original model can also be written in this form by multiplying both sides by $(1 - \phi(L))$ we have:

$$Y_t = (1 - \phi(L)) Z_t + \phi(L) Y_{t-1} + \sigma V_t. \tag{1}$$

Expanding out the term $(1 - \phi(L))Z_t$ using the lag operator we see that the two state Markov chain is transformed into a tightly parameterized 32 state chain.

A slightly different model is produced by allowing all of the parameters to switch with the Markov chain:

$$Y_t = \alpha^{s(t)} + \phi^{s(t)}(L) Y_{t-1} + \sigma^{s(t)} V_t \tag{2}$$

Three approaches to estimating the model have been taken.[3] In Hamilton's original article he developed a nonlinear filter to evaluate the likelihood function of the model and then directly maximized the likelihood function. Hamilton (1990) constructed an EM algorithm that is particularly useful for the case where all the parameters switch. Finally, Albert and Chib (1993) developed a Bayesian approach to estimation that was later refined using

[3] Links to software for all three types of estimation can be found at http://weber.ucsd.edu/~jhamilto/software.htm#Markov

results due to Chang-Jin Kim. The recent monograph by Kim and Nelson (1999) contains an excellent discussion of both Classical and Bayesian estimation of Markov switching models. The idea behind all three approaches can be illustrated in the following simple model:

$$Y_t = \alpha_0(1 - S_t) + \alpha_1 S_t + V_t, t = 1, 2, \ldots T.$$

For simplicity suppose we know that $S_0 = 1$, then entering the next period the $P[S_1 = 1|S_0 = 1] = p$. The observation Y_1 is either drawn from a normal distribution with mean α_0 and variance 1 or a normal distribution with mean α_1 and variance 1 and the likelihood is given by

$$\begin{aligned} f(y_1; \alpha_0, \alpha_1, p, q, s_0 &= 1) = \\ & \frac{p \exp(-0.5(y_1 - \alpha_1)^2) + (1-p) \exp(-0.5(y_1 - \alpha_0)^2)}{\sqrt{2\pi}} \end{aligned}$$

Assume for the moment the two mean parameters are known, then given the realization of Y_1 Bayes rule can be used to update the probability $S_1 = 1$, denote this by b_1 :

$$P[S_1|Y_1, S_0 = 1, \alpha_0, \alpha_1, p, q] = b_1$$
$$= \frac{\exp(-0.5(y_1 - \alpha_1)^2)p}{\exp(-0.5(y_1 - \alpha_1)^2)p + \exp(-0.5(y_1 - \alpha_0)^2)(1-p)}.$$

Now b_1 can be used to produce a prediction of the state next period denote this by $\widehat{b}_2$:

$$P[S_2|Y_1, S_0 = 1, \alpha_0, \alpha_1, p, q] = \widehat{b}_2 = pb_1 + (1-q)(1-b_1).$$

Using this prediction we can weight together the two possible likelihood functions depending on the state to produce a likelihood function that averages out over the value of S_2:

$$\begin{aligned} f(y_2|y_1; \alpha_0, \alpha_1, p, q, s_0 &= 1) = \\ & \frac{\widehat{b}_2 \exp(-0.5(y_2 - \alpha_1)^2) + (1 - \widehat{b}_2) \exp(-0.5(y_2 - \alpha_0)^2)}{\sqrt{2\pi}} \end{aligned}$$

and so on. This process then continues up through the last observation T to obtain the overall the likelihood function:

$$f(y_1, \ldots, y_T; \alpha_0, \alpha_1, p, q, s_0 = 1) =$$
$$\prod_{t=1}^{T} \frac{\widehat{b}_t \exp(-0.5(y_t - \alpha_1)^2) + (1 - \widehat{b}_t) \exp(-0.5(y_t - \alpha_0)^2)}{\sqrt{2\pi}}$$

Numerical optimization techniques can be used to find the maximum with respect to α_0, α_1, p, q. Further, one can also treat the probability that initial state $s_0 = 1$ as a parameter to be estimated.

For both the EM algorithm and the Bayesian analysis and inference about the unobserved Markov chain one needs to "smooth" the estimate of the Markov state b_t so that it contains information from the whole sample. Only the last probability b_T contains information on the whole observed time series. Using the Markov property and the exogeneity of the Markov chain we know that conditional on observing tomorrow's state s_{t+1}all future realizations of the observed time series $\{Y_s : s > t\}$ are not relevant for the estimate of today's state. Using this restriction and ignoring the dependence on estimated parameters we have:

$$\begin{aligned} & P[S_{T-1} = 1, S_T = 1 | Y^T] \\ &= P[S_{T-1} = 1 | S_T = 1, Y^T] P[S_T = 1 | Y^T] \\ &= P[S_{T-1} = 1 | S_T = 1, Y^{T-1}] b_T \\ &= p \frac{b_{T-1}}{\widehat{b}_T} b_T, \end{aligned} \tag{3}$$

since

$$\begin{aligned} & P[S_{T-1} = 1 | S_T = 1, Y^{T-1}] \\ &= \frac{P[S_{T-1} = 1, S_T = 1 | Y^{T-1}]}{P[S_T = 1 | Y^{T-1}]} \\ &= \frac{P[S_{T-1} | Y^{T-1}] P[S_T = 1 | S_{T-1} = 1]}{\widehat{b}_T} \\ &= \frac{b_{T-1} p}{\widehat{b}_T}. \end{aligned} \tag{4}$$

Thus denoting the smoothed probability by $\widetilde{b}_{T-1}$ we need to average out the value of S_T in (3) by performing a similar calculation for the case that $S_T = 0$:

$$\widetilde{b}_{T-1} = b_{T-1} \left(p \frac{\widetilde{b}_T}{\widehat{b}_T} + (1-q) \frac{1 - \widetilde{b}_T}{1 - \widehat{b}_T} \right).$$

This process continues until we arrive at time 1 or 0 depending on the assumption made on the initial condition. In the EM algorithm the smoothed probabilities are used to produce estimates of the unknown parameters as follows:

$$\widehat{\alpha}_0^i = \sum_{t=1}^{T} y_t(1-\widetilde{b}_{t+1}),\ \widehat{\alpha}_0^i = \sum_{t=1}^{T} y_t\widetilde{b}_{t+1}$$

and using (3):

$$\begin{aligned} \widehat{p}^i &= \frac{\sum_{t=1}^{T} P[S_{t-1}=1, S_t=1|Y^T]}{\sum_{t=1}^{T} P[S_{t-1}=1|Y^T]} = \widehat{p}^{i-1}\frac{b_{t-1}}{\widehat{b}_t}\frac{\widetilde{b}_t}{\widetilde{b}_{t-1}}, \\ \widehat{q}^i &= \frac{\sum_{t=1}^{T} P[S_{t-1}=0, S_t=0|Y^T]}{\sum_{t=1}^{T} P[S_{t-1}=0|Y^T]} = \widehat{q}^{i-1}\frac{(1-b_{t-1})}{(1-\widehat{b}_t)}\frac{(1-\widetilde{b}_t)}{(1-\widetilde{b}_{t-1})}. \end{aligned}$$

Thus, the intercepts are estimated by weighting observations by the likelihood they are in regime 0 or 1 and the transition probabilities by a pseudo count of the number of times the Markov chain stayed in the same state.

These updated parameter values are then used to re-run the filter and smoother on the observed data. This produces new parameter values and the iteration continues until a fixed point is achieved. This fixed point will be a local maxima of the likelihood function. By considering different starting parameter values for the algorithm one can check which of one local maxima is the global one. Of course one local maximum occurs at the least squares estimate of the mean of the time series with no transitions amongst states and is easy to discard (but we shall see later it causes inference problems). Consider the case of the EM algorithm where all the smoothed probabilities were 1 or 0 and there were some transitions amongst states. Then in this case the smoother would have correctly identified the movement of the Markov chain and the data would be grouped in the appropriate manner. The Bayesian approach works off this intuition by using the filter probabilities to generate realizations of the Markov chain.

Starting from b_T a value of s_T is drawn using standard inversion techniques (that is, generate a uniform random number, if it is less than or equal to $b_T, s_T = 1$, otherwise $s_T = 0$). Then using (4) if $S_T = 1$ is drawn (or its obvious complement if $S_T = 0$ is drawn), a value of S_{T-1} is drawn and this process continues until the whole sequence of the realization of the Markov chain has been drawn, $\{s_t^i\}$. Using this sequence of values for the Markov chain it is possible to split the observed data directly into two regimes. That is,

$$\begin{aligned}
\widehat{\alpha}_0^i &= \frac{1}{T_0^i}\sum_{t=1}^{T} y_t(1-s_t^i),\ T_0^i=\sum_{t=1}^{T}(1-s_t^i),\\
\widehat{\alpha}_1^i &= \frac{1}{T_1^i}\sum_{t=1}^{T} y_t s_t^i,\ T_1^i=\sum_{t=1}^{T} s_t^i.
\end{aligned}$$

In the case that flat independent priors are used for α_0, α_1,their posterior densities are normal with means $\widehat{\alpha}_0^i, \widehat{\alpha}_1^i$and variances $1/T_0^i, 1/T_1^i$ respectively. These posterior distributions can then be used to draw realizations of α_0, α_1.

The posterior of the transition parameters can also be simply found under the assumption that the priors are independent Beta distributions:

$$f(p) \propto p^{\delta_1-1}(1-p)^{\delta_2-1},\ f(q) \propto q^{\eta_1-1}(1-q)^{\eta_2-1},$$

where $\delta_1, \delta_2, \eta_1, \eta_2$ are all strictly positive and a standard uniform prior is obtained for $\delta_1 = \delta_2 = \eta_1 = \eta_2 = 1$. Given the Beta prior the posterior will also have a Beta form with parameters:

$$\begin{aligned}
\delta_1^i &= \delta_1 + \widehat{p}^i \sum_{t=1}^{T} P[S_{t-1}=1|Y^T],\ \delta_2^i = \delta_2 + (1-\widehat{p}^i)\sum_{t=1}^{T} P[S_{t-1}=1|Y^T]\\
\eta_1^i &= \eta_1 + \widehat{q}^i \sum_{t=2}^{T} P[S_{t-1}=0|Y^T],\ \eta_2^i = \eta_2 + (1-\widehat{q}^i)\sum_{t=2}^{T} P[S_{t-1}=0|Y^T],
\end{aligned}$$

where $\widehat{p}^i, \widehat{q}^i$ are defined in the same manner as in the EM algorithm. Again it is easy to draw from these posterior densities to obtain the realization p^i, q^i. Combined with the realization for the intercepts the new values can be used to run the nonlinear filter and smoother again and obtain a fresh draw of the Markov chain.

This is a Gibbs sampling algorithm for the Markov switching model. One important issue is how to initialize the algorithm. A good choice is the maximum likelihood estimates. Unlike the EM algorithm the output of the Gibbs sampler is a collection of draws from the posterior of the Markov switching model. In addition to draws of the parameters this also includes the draws of the full sample smoother. Features of the posterior distribution can then be found by averaging. For example, the estimate of the path of the Markov chain would be obtained by averaging the draws of the full sample smoother:

$$P[S_t = 1|Y^T] = \frac{1}{I}\sum_{i=1}^{I} \widetilde{b}_t^i.$$

2.2 Threshold Models

Threshold autoregressive (TAR) models are perhaps the simplest generalization of linear autoregressions. They were introduced to the time series literature by Howell Tong (see his 1983 and 1990 monographs for descriptions and extensive background on nonlinear time series).[4] Various different threshold models have been successfully applied to US GDP/GNP by Beaudry and Koop (1993), Potter (1995) and Potter and Pesaran (1997). The general form is as follows

$$Y_t = \alpha^{j(t)} + \phi^{j(t)}(L)Y_{t-1} + \sigma^{j(t)}V_t,$$

where $j(t) = 1$ if $Y_{t-d} < r_1$, $j(t) = 2$ if $r_1 \leq Y_{t-d} < r_2, \ldots$, $j(t) = J$ if $r_{J-1} \leq Y_{t-d}$, and it is possible that the length of the autoregression varies across the regimes. The parameters r_j are called the thresholds and d is called the delay.

Although this model looks very similar to the Markov switching one in (2) there is a crucial difference. In the threshold model regimes are defined by the past values of the time series itself, in the Markov switching case regimes are defined by the exogenous state of the Markov chain. Filardo (1994) constructed an intermediate case where the transition probabilities of the Markov chain vary with the history of the observed time series

Suppose $\{r_j\}, \{p_j\}, d$ were known then the model can be estimated by separating the data into groups by regime and finding the least squares estimates for the parameters in each regime. Unfortunately these parameters are not known and standard nonlinear least squares algorithms are not useful since the sum of squares functions is not differentiable with respect to these parameters. For the discrete parameters of the delay and order of autoregressive lags it is easy to repeat the least squares estimation for each choice. In the case of the threshold parameters one needs to estimate the sum of squares for a finite number of choices.

Estimation can be illustrate in a similar simplified model to the Markov switching:

$$Y_t = \alpha_0 1(Y_{t-1} < r) + \alpha_1 1(Y_{t-1} \geq r) + V_t,$$

where $1(Y_{t-1} < r) = 1$ if the inequality holds and is zero otherwise.

[4] Software for classical estimation and inference of threshold autoregressions can be obtained from www.ssc.wisc.edu/~bhansen. Software for Bayesian estimation and inference can be obtained from http://emlab.Berkeley.EDU/Software/abstracts/potter0898.html.

These latter algorithms make heavy use of recursive least algorithms that can produce big speed ups in estimation time.

The first step is to organize data into the matrix:

$$\begin{bmatrix} Y_1 & Y_0 \\ Y_2 & Y_1 \\ \vdots & \vdots \\ Y_T & Y_{T-1} \end{bmatrix},$$

then sort this matrix from smallest to largest according to the values in the second column:

$$\begin{bmatrix} Y_t^{\{1\}} & Y_{t-1}^{\{1\}} \\ Y_t^{\{2\}} & Y_{t-1}^{\{2\}} \\ \vdots & \vdots \\ Y_t^{\{T\}} & Y_{t-1}^{\{T\}} \end{bmatrix}.$$

Now note that if $r < Y_{t-1}^{\{1\}}$ then all the data are in regime 1 ($\widehat{\alpha}_1$ would be the sample average, α_0 is not identified) or if $r \geq Y_{t-1}^{\{T\}}$ then all of the data are in regime 0 ($\widehat{\alpha}_0$ would be the sample average, α_1 is not identified). If $Y_{t-1}^{\{1\}} \leq r < Y_{t-1}^{\{2\}}$ then

$$\widehat{\alpha}_0^{\{1\}} = Y_t^{\{1\}},\ \widehat{\alpha}_1^{\{1\}} = \frac{1}{T-1}\sum_{s=2}^{T} Y_t^{\{s\}},$$

and we have the (least squares) recursion:

$$\begin{aligned} \widehat{\alpha}_0^{\{i+1\}} &= \frac{T_0^{\{i\}}}{T_0^{\{i+1\}}}\widehat{\alpha}_0^{\{i\}} + \frac{1}{T_0^{\{i+1\}}}Y_t^{\{i+1\}}, T_0^{\{i+1\}} = i+1 \\ \widehat{\alpha}_i^{\{i+1\}} &= \frac{T_1^{\{i\}}}{T_1^{\{i+1\}}}\widehat{\alpha}_0^{\{i\}} - \frac{1}{T_1^{\{i+1\}}}Y_t^{\{i+1\}}, T_1^{\{i+1\}} = T-i-1 \end{aligned}$$

These estimates give the sum of squared errors function in terms of the threshold:

$$SSE_0^{\{i\}} = \sum_{s=1}^{T_0^{\{\mathtt{i}\}}} (Y_t^{\{s\}} - \widehat{\alpha}_0^{\{i\}})^2,\ SSE_1^{\{i\}} = \sum_{s=T_0^{\{\mathtt{i}\}}+1}^{T} (Y_t^{\{s\}} - \widehat{\alpha}_1^{\{i\}})^2$$

The obvious least squares estimate of r is the interval associated with the smallest sum of square errors (this is also the case when the delay is estimated). Note that any estimate within this interval is equally valid. Under the assumption of Gaussianity of the errors this would also be the maximum likelihood estimate. Chan (1993) showed that the estimate of the threshold (and delay) converges at a sufficiently fast rate that conditioning on the least squares/maximum likelihood estimate of the threshold, one could ignore its sampling variability in the asymptotic inference about the other parameters.

The Bayesian approach to estimating the threshold model (under the assumption of Gaussian errors) is similar in terms of the regime coefficients. Continuing with our simple example and assuming flat independent priors on the two intercepts, conditional on the threshold, the intercepts would have Normal distributions centered at the least squares estimates $\widehat{\alpha}_0^{\{i\}}, \widehat{\alpha}_1^{\{i\}}$with variances $1/T_0^{\{i\}}, 1/T_1^{\{i\}}$ respectively. In order to find the marginal posterior of the threshold consider the case of a flat prior on the thresholds. Then the joint posterior is proportional to the likelihood function.

$$\begin{aligned} p(\alpha_0, \alpha_1, r|Y^T) \quad &\propto \\ & \frac{1}{\sqrt{2\pi}^{T_0^{\{i\}}}} \exp(-0.5 \sum_{s=1}^{T_0^{\{i\}}} (Y_t^{\{s\}} - \alpha_0)^2) \\ & \times \frac{1}{\sqrt{2\pi}^{T_1^{\{i\}}}} \exp(-0.5 \sum_{s=T_0^{\{i\}}+1}^{T} (Y_t^{\{s\}} - \alpha_1)^2). \end{aligned}$$

Subtracting and adding the least squares estimates of the intercepts in each squared function and re-arranging using the orthogonality of least squares estimates we have:

$$\begin{aligned} p(\alpha_0, \alpha_1, r|Y^T) \quad &\propto \\ & \exp\left(-0.5\left[\sum_{s=1}^{T_0^{\{i\}}} (Y_t^{\{s\}} - \widehat{\alpha}_0^{\{i\}})^2 + \sum_{s=T_0^{\{i\}}+1}^{T} (Y_t^{\{s\}} - \widehat{\alpha}_1^{\{i\}})^2\right]\right) \\ & \times \exp\left(-0.5\left[T_0^{\{i\}}(\widehat{\alpha}_0^{\{i\}} - \alpha_0)^2 + T_1^{\{i\}}(\widehat{\alpha}_1^{\{i\}} - \alpha_1)^2\right]\right). \end{aligned}$$

Now integrating out over α_0, α_1 we have:

$$\begin{aligned} p(r|Y^T) \quad &\propto \\ & \frac{1}{\sqrt{T_0^{\{i\}} T_1^{\{i\}}}} \exp\left(-0.5\left[SSE_0^i + SSE_1^i\right]\right)(Y_{t-1}^{\{i\}} - Y_{t-1}^{\{i-1\}}), \quad (5) \end{aligned}$$

for $i = 2, \ldots, T$. Where the intervals $(Y_{t-1}^{\{i\}} - Y_{t-1}^{\{i-1\}})$ represent the fact that there is no information on the threshold between data points.

The Bayesian modal estimate of the threshold is not likely to be the same as the classical one (i.e., the same interval) even under the assumption of a flat prior on the thresholds. Although the exponential term will be maximized at the same threshold value as the total sum of square errors, the lead term involving the inverse of the square roots of the sample size in each regime will affect the location of the maximum. Further, the mean and median of the posterior distribution of the threshold are very unlikely to be at the mode unless T is very large and the sum of squared errors terms dominates.

Marginal inference about the regime coefficients is very different than in the classical case (see Hansen 1999 for a method to improve the classical approach). In the Bayesian case inference about the intercepts would be based on weighting the individual normal distributions by the posterior probability of the particular threshold interval and threshold uncertainty would affect inference about individual regime coefficients.

2.3 Smooth Transition Autogressions

For many the abrupt regime changes in the threshold model are unrealistic. Consider the case where one is forecasting US GDP and the initial release his slightly below the threshold but the subsequent revision is above the threshold. A threshold model would imply large changes in the forecast of the future for this small change initial conditions. Further, the difficulties of the non-standard likelihood/least squares functions are a distraction.

As originally suggested by Chan and Tong (1986) and subsequently developed by Timo Teräsvirta and his various co-authors (see for example, his 1993 monograph with Clive Granger) if one introduces smooth transitions between regimes standard nonlinear estimation techniques can be used. Since smooth transition models have a more traditional structure, Teräsvirta has been able to implement a model specification, estimation and diagnostic cycle very similar to the Box and Jenkins approach (see his 1994 JASA paper). The models were successfully applied to a wide range of industrial production series by Teräsvirta and Anderson (1992).

In the simple threshold model above imagine changing from the indicator function to a smooth cumulative distribution function:

$$Y_t = \alpha_0(1 - F(Y_{t-1}; \gamma, r)) + \alpha_1 F(Y_{t-1}; \gamma, r) + V_t,$$

where $F(-\infty; \gamma, r) = 0,\ F(\infty; \gamma, r) = 1.$

The simplest smooth transition function is of a logistic type:

$$F(Y_{t-1};\gamma,r)=\frac{1}{1+\exp(-\gamma(Y_{t-1}-r))},$$

where the parameter $\gamma>0$ determines the abruptness of the transition at r. For example, for very large γ the smooth transition model might effectively be the same as a threshold model for certain values of r since for the pair observations of Y_{t-1} either side of r the transition might be complete. On the other hand, if $\gamma\simeq 0$ then the logistic function hardly varies away from 0.5 and there is really only one regime.

Once again if r and γ were known ex-ante simple least squares methods could be used to estimate the remaining parameters. Thus, one can concentrate the least squares function/likelihood function with respect to r and γ and use standard nonlinear optimizers to estimate these parameters. In practice because of numerical instability issues it makes some sense to limit the variation in one of these two parameters. One choice favored by Teräsvirta is to normalize $(Y_{t-1}-r)$ by the standard deviation of the delay variable. Another is to examine a finite set of thresholds, as in the Threshold autoregression case, thus leaving only γ to be directly estimated by the nonlinear optimizer.[5]

A more general version of the model is as follows:

$$Y_t=\alpha_1+\phi_1(L)Y_{t-1}+(\alpha_2+\phi_2(L)Y_{t-1})\,F(Y_{t-d};r,d,\gamma)+\sigma V_t \qquad (6)$$

One difference to the other two models is that less attention is focused on possible changes in the variance of the innovations across regimes. In addition to the logistic smooth transition function there has also been considerable attention paid to the possibility of symmetric transitions away from the threshold:

$$F^s(Y_{t-1};\gamma,r)=1-\exp(-\gamma(Y_{t-1}-r)^2).$$

Lubrano (1999) discusses Bayesian estimation of smooth transition autoregressions. As in the threshold autoregression case one can integrate out the intercept, autoregressive coefficients and variance. This leaves a 3-dimensional posterior distribution. One dimension is the delay parameter and is discrete, the other two involve the threshold and speed of transition. One choice is to find their posterior using standard numerical techniques. Another choice is to use a Metropolis-Hastings algorithm.

[5] Smooth transition autoregressions can be estimated using standard econometric packages with nonlinear estimation options. In addition there will be a website available soon with Gauss software available implementating the full approach taken by Teräsvirta.

3 TESTING

Perhaps the greatest theoretical progress in the last ten years has been in our understanding of testing for nonlinearity in economic time series. On the other hand, perhaps the least empirical progress has been made in finding evidence for nonlinearity in economic time series given the new theoretical tools available. For example, a range of statistical tests have been applied to aggregate output in the United States. At first the results were encouraging but as shown by Hansen (1992,1996a) some of the encouragement was of the wishful thinking variety. First, direct tests of the Hamilton's Markov switching model suggested that it was not statistically significant. Subsequent searches over different Markov switching specifications (Chib 1995, Hansen 1992) were motivated by this failure and their success should be qualified. Second, direct tests of threshold models also indicate that the nonlinear terms are not highly statistically significant. Tests for smooth transition models usually do not take into account the effects of searching over different values of the delay parameter. These classical statistical results have also been supported by the Bayesian analysis of Koop and Potter (1999a). Overall there is probably less evidence for nonlinearity in US output at the end of the 1990s then researchers thought at the start of the decade but still considerable evidence that the behavior of output in at business cycle turning points is not well captured by linear models.[6]

In Hamilton's (1989) paper he was careful to point out that testing the null hypothesis of a linear model against his particular nonlinear model was not standard and appeared to be very difficult. There were two main problems. First, under the null hypothesis the transition parameters of the Markov chain in the alternative hypothesis were not pinned down. Consider the likelihood value from above:

$$\begin{aligned} f(y_1; \alpha_0, \alpha_1, p, q, s_0 &= 1) = \\ & \frac{p\exp(-0.5(y_1-\alpha_1)^2)+(1-p)\exp(-0.5(y_1-\alpha_0)^2)}{\sqrt{2\pi}} \end{aligned}$$

In the case that $\alpha_1 = \alpha_0$ the value of p has no effect on the likelihood function.

Second, the likelihood function for the nonlinear model has a local maximum at the parameter values for the linear model with p set at the boundary value of 1 since this implies all the realizations of the Markov chain will be in state 1 given that this is the initial condition.

[6] For a review of earlier testing approaches see Brock and Potter (1993).

The first problem is common to nearly all nonlinear models with the notable exception of the smooth transition class where the following reparametrization is available:

$$F^*(Y_{t-1}; r, \gamma) = \frac{1}{1 + \exp(-\gamma(Y_{t-d} - r))} - 0.5.$$

Now testing for nonlinearity can be undertaken by allowing γ to take on all real values under the alternative and fixing the delay at a particular value. The null hypothesis of linearity is captured by the restriction $\gamma = 0$. As described in Teräsvirta (1994) a Lagrange Multiplier test can be developed. However, in the general case where the delay is unknown the problem crops up again.

In the general case the imposition of the null hypothesis of linearity leaves some of the parameters describing the nonlinear model free. In statistics it is called the Davies' problem. In order to understand the seriousness of the Davies' problem consider that under the null hypothesis for a fixed choice of the parameters the likelihood ratio will have a Chi-squared distribution in large samples. But one can vary these "free" parameters to find a largest and smallest likelihood ratio. Again under the null hypothesis these are both draws from the same Chi-squared distribution. Obviously if the minimum value still exceeds the critical value rejection of the null hypothesis is warranted. Such an outcome is possible in the case where some restrictions are placed on the parameters present only under the alternative. For example, in the threshold case if each regime must have at least 15% of the data or in the Markov switching case where one rules out boundary values for the transition probabilities and defines them on a closed subset of the unit interval. One solution is to choose the nuisance parameters at random. Another more powerful one is to examine the properties of the maximum across the parameter space as described in the accompanying article by Bruce Hansen.

All of the previous solutions ignored some information in the behavior of the test statistic as the parameters describing the nonlinear model are varied. In a large enough sample it is safe to do this and concentrate on the most powerful tests involving the largest test statistic. However, in a typical macroeconomic application where the likelihood function might have may local maxima such an approach can be dangerous. One obvious solution would be to examine some average of the test statistics. This intuition was formalized by Andrews and Ploberger (1994) who showed that under certain conditions averaged test statistics were the most powerful.

In order to illustrate these ideas I will work with the simple threshold autoregression as the alternative nonlinear model.

$$Y_t = \alpha_0 1(Y_{t-1} < r) + \alpha_1 1(Y_{t-1} \geq r) + V_t$$

and the sequence of observed data $y_0 = -0.2, y_1 = 2, y_2 = -0.1, y_3 = -1.9$. The initial observation will be treated as fixed and for both linear and nonlinear models the innovation variance is assumed to be 1. For the linear model we have a sample average of 0 and a resulting sum of squared errors of $2^2 + 0.1^2 + 1.9^2 = 7.62$.

For the threshold model we shall assume a priori that the threshold value is in the interval $[-1, 1]$. Thus we have 3 possible sum of squared error functions:

1. If $-1 \leq r \leq -0.2$, then the same sum of squared error function is the sum as the linear model since all the observations are drawn from the same regime.

2. If $-0.2 < r \leq -0.1$ then observations 2 and 3 are drawn from the upper regime and observation 1 is drawn from the lower regime. Thus, $\widehat{\alpha}_0 = 2, \widehat{\alpha}_1 = (-0.1 + -1.9)/2 = -1$ and the sum of squared errors is $0^2 + 0.9^2 + 0.9^2 = 1.62$.

3. If $-0.1 < r \leq 1$ then observations 1 and 3 are drawn from the lower regime and observation 2 is drawn from the upper regime. Thus, $\widehat{\alpha}_0 = (2 + -1.9)/2 = 0.05, \widehat{\alpha}_1 = -0.1$ and the sum of squared errors is $1.95^2 + 0^2 + 1.95^2 = 7.605$.

We have three log likelihood ratio statistics: $0, 6, 0.03$ (in this example the log likelihood ratio is just the difference in the sum of squared errors). Clearly, an estimate of r in the interval $-0.2 < r \leq -0.1$ is the maximum likelihood estimate and the associated test statistic of 6 is large relative to a Chi-squared distribution with one degree of freedom that one would use to measure statistical significance in large samples. On the other hand the minimum statistic of 0 is obviously not significant compared to a Chi-squared distribution. If we average the test statistics against a uniform distribution for the threshold in the given interval the value is $0.4 \times 0 + 0.05 \times 6 + 0.55 \times 0.03 = 0.317$. This is also the expected value of a randomized test but such a test would have considerable variability depending on which interval was chosen.

Hansen (1996a) provides a general method of calculating sampling distributions under the null hypothesis for such operations on the family of test statistics. The procedure in this case works as follows (note in this special case the procedure gives an exact small sample result, this is not true in general). Generate 3 standard normal random variables and without loss of generality assume they are ordered: $V_1 \leq V_2 \leq V_3$. The likelihood ratio statistic considered by plugging in the maximum likelihood estimate of the threshold is equivalent to the following minimization problem:

$$\frac{4}{9}(V_1^2 + V_2^2 + V_3^2) - \frac{1}{4}\min\{V_1^2 + V_2^2, V_2^2 + V_3^2\}$$

The 95th percentile of this distribution is approximately 2.9 thus on this measure there is significant evidence of the threshold effect (the p-value for the observed statistic is 0.1%). This give significance at a better than 1% level. However, if we consider the average likelihood ratio statistic using the observed frequency of likelihood ratios different than zero:

$$0.6\left[\frac{4}{9}(V_1^2 + V_2^2 + V_3^2) - \frac{1}{4}(V_1^2 + V_2^2) - \frac{1}{4}(V_2^2 + V_3^2)\right],$$

the 95th percentile is approximately 1.4. Thus, on this measure the threshold effect is not statistically significant.

An alternative to the classical approach is to use Bayes factors to compare the linear and nonlinear models (see Koop and Potter, 1999a). The Bayes factor in this case is ratio of the average value (over prior distribution of parameters) of the likelihood function for the nonlinear model against the average value (over the prior distribution of parameters) of the likelihood function for the linear model. In order for the Bayes factor to be useful, informative priors on the parameters are required. In this simple case we shall assume that the prior on the mean for the linear model is uniform over $[-m, m]$ and that the prior on the nonlinear model is uniform and independent for both intercepts over the same interval of length $2m$. Then using our previous result on the posterior for the threshold and the numbers for differences in sum of squared errors form above we have:

$$\begin{aligned} & \frac{\sqrt{2\pi}}{2m}\left[0.4 \times 1 + 0.05 \times \frac{\sqrt{3}}{\sqrt{2}}\exp(3) + 0.55 \times \frac{\sqrt{3}}{\sqrt{2}}\exp(0.015)\right] \\ = & \frac{1.2533}{m}2.314 \text{ if } m \geq 2. \end{aligned}$$

Here to keep things simple I have assumed that the prior is sufficiently flat that boundary value problems do not occur. The interpretation of the Bayes factor is different to classical statistical tests. In this case it is giving (assuming linear and nonlinear models were equally likely ex-ante) the posterior odds in favor of the nonlinear model. Notice that if $2 \leq m < 1.253 \times 2.314 = 2.899$ they are favorable. For an ignorant prior ($m \to \infty$) the Bayes factor would always favor the simpler linear model. The highest posterior odds for the nonlinear model are obtained at $m = 2$ and are equal to about 1.45, that is about 60% of the posterior weight is placed on the nonlinear model and about 40% on the linear model.

Obviously the main drawback of the Bayesian approach is its sensitivity to the prior distributions on the parameters. For example, if the prior on the threshold had been uniform on $[-10, 10]$ then the Bayes factor would be

$$\begin{aligned} & \frac{\sqrt{2\pi}}{2m}\left[\frac{9.8}{20} \times 1 + \frac{0.1}{20} \times \frac{\sqrt{3}}{\sqrt{2}}\exp(3) + \frac{10.1}{20} \times \frac{\sqrt{3}}{\sqrt{2}}\exp(0.015)\right] \\ = & \frac{1.253}{m}1.241 \text{ if } m \geq 2, \end{aligned}$$

and at best ($m = 2$) the posterior odds in favor of nonlinearity would be 44%. The benefit of the drawback is that the Bayes factor has a very big penalty for more complicated models something that classical models are unable to do. For example, in the case that both the threshold and intercepts are assumed to uniformly distributed over a wide interval we are placing most of the ex-ante weight on very strong forms of nonlinearity. Suppose $m = 10$ and $r \sim U[-10, 10]$ then most threshold models produced would have bimodal distributions with modes very far apart and values like $y_2 = -0.1$ are improbable.

Bayesian methods have a strong advantage when it comes to finding tests for Markov switching models. Hansen (1992,1996b) provides a bounds test that deals with the both the Davies' problem and the zero scores problem. However, this is a very computationally intensive test and it only provides a bound on the size. Garcia (1998) adopts the approach of Andrews and Ploberger (1994)and Hansen (1996a) by ignoring the zero scores problem. His simulation results suggest that zero scores might not be a problem in practice. In contrast, Bayesian testing of the Markov switching models is simple and direct. In order to illustrate consider the simple Markov switching model:

$$Y_t = \alpha_1 S_t + V_t.$$

Here linearity is given by the restriction that $\alpha_1 = 0$. As discussed in Koop and Potter (1999a) the Bayes factor in this Markov switching case can be written as the ratio of the posterior density to the prior density for α_1 evaluated at zero, i.e.,

$$\frac{b(\alpha_1 = 0)}{p(\alpha_1 = 0|Y^T)}.$$

Once again suppose that α_1 has a uniform prior on the interval $[-m, m]$ and using our previous results we have a conditional Bayes factor of :

$$\frac{b(\alpha_1 = 0)}{p(\alpha_1 = 0|Y^T, \{s_t^i\}, p^i, q^i)} = \frac{\sqrt{T_1^i}}{2m\sqrt{2\pi}} \exp\left[-0.5T_1^i(\widehat{\alpha}_1^i)^2\right].$$

The overall Bayes factor is found by replacing the conditional posterior density with its average across posterior draws:

$$p(\alpha_1 = 0|Y^T) = \frac{1}{I}\sum_{i=1}^{I} \frac{\sqrt{T_1^i}}{\sqrt{2\pi}} \exp\left[-0.5T_1^i(\widehat{\alpha}_1^i)^2\right],$$

which requires minimal changes to existing computer code.

Once again if the initial ignorance about α_1 is high (large values of m) there is less chance of finding evidence of nonlinearity. Or alternatively if the priors on p and q lead to most runs of $\{s_t^i\}$ consisting of zeros there is also little chance of finding nonlinearity. Chib (1995) discusses a slightly more computationally intensive method of constructing the average likelihood for the Markov switching model that is easier to implement when all the parameters of the model switch with the Markov variable as in (2).

4 CONSTRUCTING CONDITIONAL EXPECTATIONS

Once a final nonlinear model is arrived at there remains the issue of understanding the estimated dynamics and forecasting capabilities. Since the primary objective of nonlinear modelling is to obtain the true conditional expectation function there is still a substantial task remaining.

The task is easiest for the Markov switching models. As an illustration consider Hamilton's original model in the form given in (1). We introduce the following notation: P represents the transition matrix of the 32 state Markov chain, b_t is a vector representing the filter probabilities for each of the 32 individual states, s^* is a vector containing the 32 possible values of $(1-\phi(L))Z_t$.

$$E_t[Y_{t+h}] = E_t[Z^*_{t+h}] + E_t[\phi(L)Y_{t+h-1}].$$

The second term on the LHS can be evaluated using standard linear recursions. The first term on the LHS requires more care. Assume for the moment that S^*_t was in the information set at time t. Then one could find $E_t[S^*_{t+h}]$ using the estimated probability transition matrix as follows:

$$E[S^*_t|S^*_t = s_j] = \mathsf{s}^{*'}\mathsf{P}^{h\prime}\mathsf{e}_j,$$

where e_j is a vector of zeros except for the jth row which contains 1. Of course the state of the Markov chain is not known at time t but one can replace e_j with filter probabilities over the states of the Markov chain at time t,

$$E[S^*_t|S^*_t = s_j] = s^{*'}\mathsf{P}^{h\prime}\mathsf{b}_t.$$

For the TAR and STAR models obtaining the conditional expectation nearly always requires the use of simulation once the forecast horizon exceeds the length of the delay lag. If the horizon is less than the delay lag then the conditional expectation is given by iterating on the nonlinear difference equation. For example, in the simple model:

$$Y_t = -1(Y_{t-4} < 0) + 1(Y_{t-4} \geq 0) + V_t,$$

the conditional expectation function up to horizon 4 is given by

$$E[Y_{t+h}|Y^t] = -1(Y_{t+h-4} < 0) + 1(Y_{t+h-4} \geq 0), \text{ if } h \leq 4.$$

Suppose the conditional expectation of Y_{t+1} at time t was 1 (i.e., $Y_{t-3} > 0$). It would be incorrect to use this value in the indicator function to forecast $Y_{t+5} = 1$ since V_{t+1} is standard normal there is non-zero probability that $\mathrm{Y}_{t+1} < 0$. Continuing with this example, there is approximately a 16% probability that $\mathrm{Y}_{t+1} < 0$. Thus,

$$\begin{aligned} E[Y_{t+5}|Y_{t-3} &> 0] = -P[Y_{t+1} < 0|Y_{t-3} > 0] + P[Y_{t+1} \geq 0|Y_{t-3} > 0] \\ &= -0.16 + 0.84 = 0.68. \end{aligned}$$

Similar results would be available up to forecast horizon 8 where things become more complicated:

$$E[Y_{t+9}|Y_{t-3} > 0] = -P[Y_{t+5} < 0|Y_{t-3} > 0] + P[Y_{t+5} \geq 0|Y_{t-3} > 0].$$

In order to evaluate the probabilities on the RHS one can iterate forward the distribution for Y_{t+5} using the fact that Y_{t+1} is drawn from a $N(1,1)$ with probability 0.84 and from a $N(-1,1)$ with probability 0.16.

Notice that forecasts after horizon 4 were crucially dependent on the assumption on the innovations were Gaussian and the size of their variance, unlike the linear case. Further the calculations were greatly simplified by the fact there were no autoregressive lags. Consider another simple model:

$$Y_t = 0.5 Y_{t-1} 1(Y_{t-1} \geq 0) + V_t.$$

We have $E[Y_{t+1}|Y^t] = 0.5 Y_t 1(Y_t \geq 0)$ and at horizon 2

$$E[Y_{t+2}|Y^t] = 0.5 E[Y_{t+1} 1(Y_{t+1} \geq 0)|Y^t],$$

which can be evaluated using the fact the $Y_{t+1} 1(Y_{t+1} \geq 0)$ is a (conditional) truncated normal with parameters $0.5 Y_t 1(Y_t \geq 0)$ and 1 to obtain

$$E[Y_{t+2}|Y^t] = 0.25 Y_t 1(Y_t \geq 0) + 0.5 \left[\frac{\varphi(-0.5 Y_t 1(Y_t \geq 0))}{1 - \Psi(-0.5 Y_t 1(Y_t \geq 0))} \right],$$

where $\varphi(z)$ and $\Psi(z)$ are the standard normal density and cumulative distribution functions respectively. Notice that for $Y_t \gg 0$ the forecasts are very similar to a first order linear autoregression with zero intercept and autoregressive coefficient of 0.5. For smaller values of Y_t the forecasts are very different from such a linear autoregression. However,

$$E[Y_{t+3}|Y^t] = 0.5 E[Y_{t+2} 1(Y_{t+2} \geq 0)|Y^t],$$

is much less tractable since Y_{t+2} does not have a conditional normal distribution.

Instead of directly attempting to calculate this expectation consider using the time series model to simulate a large number of time series for each particular history. Thus, we know that $Y_{t+1} \sim N(0.5 Y_t 1(Y_t \geq 0), 1)$ and this fact can be used to generate K realizations from this distribution. Now take this realizations and generate K realizations of Y_{t+2} using K draws of V_{t+2} and the equation:

$$y_{t+2}^k = 0.5 y_{t+1}^k 1(y_{t+1}^k \geq 0) + v_{t+2}^k.$$

We can then approximate $E[Y_{t+3}|Y^t]$ by:

$$\frac{1}{K} \sum_{k=1}^{K} 0.5 y_{t+2}^k 1(y_{t+2}^k \geq 0)$$

which by the Law of Large Numbers will converge to the conditional expectation as $K \to \infty$.

For both the Threshold and Smooth Transition Models dynamic simulation of time series paths allows one to calculate good approximations to the conditional expectation function for a wide range of specifications.

4.1 Forecasting with known parameters

Given a method to calculate the conditional expectation function and a model with no unknown parameters the production of forecasts is straightforward. The comparison of these forecasts with those from linear models is less straightforward. First, in out of sample comparisons it is important that the nonlinear feature found in the historical sample is present. For example, many of the nonlinear models of U.S. output focus on the dynamics entering and recovering from recessions. The last recession in the United States ended in early 1991, thus it has been difficult to verify the nonlinearities out of sample.

This lack of variation in the out of sample period can be compensated for by various experiments within the sample. In particular, unlike linear models useful information can be generated by considering in sample multistep ahead prediction. Consider first a linear first order autoregression with zero intercept and estimated first order coefficient $\widehat{\phi}$ and residuals $\{\widehat{U}_t\}$. We have the following identities:

$$\begin{aligned} Y_{t+1} &= \widehat{\phi} Y_t + \widehat{U}_{t+1}, \\ Y_{t+2} &= \widehat{\phi} Y_{t+1} + \widehat{U}_{t+2} = \widehat{\phi}^2 Y_t + \widehat{\phi}\widehat{U}_{t+1} + \widehat{U}_{t+2}, \dots , \\ Y_{t+H} &= \widehat{\phi} Y_{t+H-1} + \widehat{U}_{t+H} = \widehat{\phi}^H Y_t + \sum_{h=0}^{H-1} \widehat{\phi}^h \widehat{U}_{t+H-h}. \end{aligned}$$

Thus, for large sample size the in sample one-step ahead forecast variance is $\widehat{\sigma}^2$,the two-step ahead is $\widehat{\sigma}^2(1+\widehat{\phi}^2)$, and h-step ahead by $\widehat{\sigma}^2(\sum_{h=0}^{H-1} \widehat{\phi}^{2h})$.

For nonlinear models such a recursion does not exist. In fact it is not even possible to show that the RMSE of a forecasts grows monotonically with the horizon as in a linear model for all histories. This cost in terms of computational complexity is a benefit in terms of model evaluation since simulation of h-step ahead conditional expectation function in sample can provide a diagnostic check on the nonlinear model. By definition, for any forecast horizon, the variance of the forecast error using the best linear predictor has to be greater than or equal to the variance of the forecast error using the true conditional expectation function, Thus, a good diagnostic check on the estimated nonlinear model is whether this is true in the observed sample.

There are two ways to formalize the notion of the best linear predictor. One is to use the linear model used in the testing phase and iterate it as above to obtain multi-step predictions. In this case it might prove useful to adjust for the loss of observations as the steps ahead of the prediction increase. For Markov switching models this is an easy check since no simulation is required. For the other models it becomes more computationally intensive once the forecast horizon exceeds the delay lag of the threshold variable.

Using the iterated properties of a linear model is only a weak check since if nonlinearity is present the model will only produce the best linear forecasts for one step ahead. A stronger one in the case where nonlinearity is present but not necessarily of the type estimated, is to estimate different linear models for each prediction horizon. If the true nonlinearity is different from that estimated then such adaptive linear models should start to outperform the nonlinear model.

4.2 Forecasting with parameter uncertainty

It is typical in linear time series forecasting applications to provide some measure of the effect on the forecast of parameter uncertainty. There are a two main ways of doing this using classical statistical methods. One can use asymptotic approximations for functions of the parameters of the model or one can simulate random draws from the approximately normal sampling distributions of the parameters and construct the forecasts from the draws. These methods are only directly applicable to the STAR model. Even in this case the situation is far more complicated than the linear one, since we need to calculate the conditional expectation function, as described above, for each set of parameter values.

In the case of the threshold model, the threshold estimates and delay are converging at faster rates than the other parameters, hence in a large enough sample they can be ignored. Unfortunately, the adequacy of this large sample approximation in time series of the typical length in macroeconomics is very much open to question. Further, unlike the mild effects that a poor approximation might have in a linear model the effects in threshold models can be drastic. Consider the case where the threshold variable in the information set for the forecast is close to the estimate of the threshold. By treating the threshold and delay as known the forecast will be very different depending on which side of the threshold the observed data is. However, this is a false precision since the value of the threshold or delay is not known with certainty.

Dacco and Satchell (1999) find that the regime misclassification introduced can lead to a nonlinear model having inferior forecast performance to a linear model even when the nonlinear model is true and all parameters but the threshold are known. This issue can be illustrated using the simple intercept shift threshold autoregression under the assumption that α_0, α_1 are known. In this case the posterior distribution for the threshold is given by:

$$p(r|Y^T) \propto \exp\left(-0.5\left[SSE_0^i + SSE_1^i\right]\right)(Y_{t-1}^{\{i\}} - Y_{t-1}^{\{i-1\}}).$$

Denoting the cumulative distribution function of the posterior of r by G we have for the one step ahead forecast:

$$E_T[Y_{T+1}] = \alpha_0 + (\alpha_1 - \alpha_0)G(y_T).$$

Obviously as the sample size increases $G(y_T)$ will start to get closer to an indicator function but in typical macroeconomic samples it will contain considerable uncertainty about the true location of the threshold.

One implication of the above analysis is that for forecasting Bayesian estimation of threshold models has a distinct advantage (although it might be possible to adapt some of results of Hansen 1999 to reproduce a classical analog). This is also true for the Markov switching models. In this case the difficulty is that the estimate of the current Markov state is dependent on the whole set of parameter estimates. Any change in the parameter estimates would affect the estimate of the Markov state. One could simulate from the asymptotic distribution of the parameter estimates and then rerun the filter at these parameter values to obtain some feeling for how the estimate of the current state would change. But this is an inconsistent approach: the new filter values would imply different parameter estimates (by definition only maxima in the likelihood function will not have this problem, see the discussion of the EM algorithm above). The Bayesian solution to the forecasting problem is to use the output of the Gibbs sampler. Recall that for each complete draw from the Gibbs sample we have draws of the unknown parameter values and the value of the filter probability of the most recent state. These values can be used to form a forecast. Repeating this exercise across all the draws from the posterior and averaging will produce the conditional expectation allowing for parameter uncertainty.

4.3 Impulse Response Functions

Once one is satisfied with the nonlinear model there is the remaining question of describing how its dynamics differ from that of linear models fit to the same time series. Since most economists describe the dynamics of linear models using impulse response functions it is important to generalize impulse response functions to nonlinear time series. As described in the introduction the linear dynamics of a time series are given by the Wold Representation. The coefficients in the Wold Representation can be thought of as producing the same answer to the following four questions:

1. What is the response to a unit impulse today when all future shocks are sent to zero?
2. What is the response to a unit impulse today when all future shocks are integrated out?
3. What is the derivative of the predictor of the future?
4. How does the forecast of the future change between today and yesterday, normalizing the change by the innovation to the time series today.

For nonlinear models there will be different answers to each question. To illustrate consider the simple threshold model:

$$Y_t = -1(Y_{t-1} < 0) + 1(Y_{t-1} \geq 0) + V_t$$

and the case of a positive unit impulse for the first two questions.

1. If $Y_t \geq 0$ then the response is 0 for all horizons , if $-1 \leq Y_t < 0$ then response is 2 for all horizons since this permanently moves the time series into the upper regime given the assumption of no future shocks, if $Y_t < -1$ then the response is 0 for all horizons.
2. At horizon 1 the response is the same as the answer to question 1 but as the horizon increases we have the difference:

   $$E[Y_{t+h}|Y_t = y_t + 1] - E[Y_{t+h}|Y_t = y_t],$$

   which must converge to zero by the stationarity of the underlying time series.
3. The derivative is either 0 or not defined if $Y_t = 0$.

4. Consider the case where $E_{t-1}[Y_t] = 1$ and the realized value of $Y_t = 1$. Then the initial shock is 0 but

$$E_t[Y_{t+1}] - E_{t-1}[Y_{t+1}] = 1 - 0.68 = 0.32.$$

Or the case where $E_{t-1}[Y_t] = 1$ and the realized value of $Y_t = 5$. Then the initial shock is 4 but the $E_t[Y_{t+1}] - E_{t-1}[Y_{t+1}]$ is still equal to 0.32.

It should be immediately apparent that questions 1 and 3 are not particularly useful questions to ask for a nonlinear time series. This leaves a choice between the more traditional definition of the impulse response function defined by the answer to question 2 and the forecasting revision function defined by the answer to question 4. In order to choose between the two possibilities observe that both the initial condition and the magnitude and sign of the impulse is important in describing the dynamics of nonlinear models. This is problematic since one can chose values of the initial condition or shock that produce atypical responses. In answering question 4 the properties of the impulse are defined directly by the time series model, that is:

$$E_t[Y_t] - E_{t-1}[Y_t],$$

or σV_t in our examples. Further, in order to define the initial conditions one can use the history of the time series or random draws from its simulated distributions. In answering question 2 there is no direct way of defining the relevant set of perturbations away from the initial condition using the properties of the time series model. Note that using the time series innovation to define the perturbation is not correct since this innovation represents the unforecastable change between time $t - 1$ and t.

Koop, Pesaran and Potter (1996) and Potter (1999) call the forecast revision function a generalized impulse response function and develop its properties. Of particular importance is the fact that the generalized impulse response function is a random variable on the same probability space as the time series. Thus, in order to measure the size of a response at a particular horizon one needs to measure the size of a random variable. In the cases where the response averages to zero (the standard one for the forecast revision function) one can use the concept of second order stochastic dominance. Some other experiments lead to responses with non-zero means where size can be measured by the mean. Perhaps the most interesting nonlinear feature found from impulse response functions has been a lower level of persistence of negative shocks in recessions than expansions (see Beaudry and Koop 1993).

In the linear time series literature there is a considerable literature on inference for impulse response functions. As the previous discussion of forecasting under parameter uncertainty would indicate such inference is more difficult for nonlinear time series models. Koop (1996) develops a Bayesian approach to the analysis of generalized impulse response functions that allows directly for parameter uncertainty. Koop and Potter (1999b) combine this analysis of impulse response functions under parameter uncertainty with Bayesian measures of model uncertainty (see the discussion at the end of Section 3) to present impulse response functions that are less reliant on particular model specifications.

5 CONCLUDING REMARKS

Three basic nonlinear time series models have been reviewed. Both Classical and Bayesian approaches to estimation and inference have been described. The focus has been on univariate time series models. The three basic models do generalize to the multiple time series case but some of the difficulties concerning inference highlighted in this article are compounded in higher dimensions. Given the difficulties of interpreting test statistics for linearity vs. nonlinearity discussed, it seems important to shift the focus to differences in forecasting and dynamics that allow for both parameter and model uncertainty.

References

[1] Albert, J. and Chib, S. (1993). “Bayesian inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts,” Journal of Business and Economic Statistics, 11, 1-15.

[2] Andrews, D. and Ploberger, W. (1994). “Optimal tests when a nuisance parameter is present only under the alternative,” Econometrica, 62, 1383-1414

[3] Beaudry, P. and Koop, G. (1993). “Do recessions permanently change output?” Journal of Monetary Economics, 31, 149-163.

[4] Brock, W., and Chamberlain G., (1984) “Spectral Analysis cannot tell a macroeconometrician whether his time series came from a stochastic

economy or a deterministic economy," *SSRI WP* 8419, University of Wisconsin Madison.

[5] Brock, W., and Potter, S., (1993) "Nonlinear time series and macroeconometrics," in G.S. Maddala et al. (eds) *Handbook of Statistics*, North Holland, Amsterdam.

[6] Chan, K.S., (1993) "Consistency and limiting distribution of the least squares estimator of a threshold autoregression," *Annals of Statistics, 21, 520-533.*

[7] Chan, K.S. and Tong, H. (1986) "On estimating thresholds in autoregressive models," *Journal of Time Series Analysis*, 7, 178-190.

[8] Chib, S. (1995). "Marginal likelihood from the Gibbs output," *Journal of the American Statistical Association*, 90, 1313-1321

[9] Dacco R., and Satchell, S., (1999) "Why do regime-switching models forecast so badly," *Journal of Forecasting*, 18, 1-16.

[10] Filardo, A., (1994) "Business cycle phases and their transitional dynamics," *Journal of Business and Economic Statistics*, 12, 299-308.

[11] Gallant, A.R., P.E. Rossi and G. Tauchen (1993). "Nonlinear Dynamic Structures," *Econometrica*, 61, 871-908.

[12] Garcia, R., (1998) "Asymptotic null distribution of the likelihood ratio test in Markov switching models," *International Economic Review*, 39, 763-788,

[13] Granger, C.W.J. (1983)"Forecasting white noise," in *Applied Time Series Analysis of Economic Data*, Proceedings of the Conference on Applied Time Series Analysis of Economic Data, edited by A. Zellner, U.S. Government Printing Office.

[14] Granger, C.W.J. and Teräsvirta, T., *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

[15] Hamilton, J. (1989). "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, 57, 357-384.

[16] Hamilton, J. (1990) "Analysis of time series subject to changes in regime," *Journal of Econometrics*, 45, 39-70.

[17] Hansen, B. (1992) "The likelihood ratio test under non-standard conditions: testing the Markov switching model," Journal of Applied Econometrics, 7, S61-S82.

[18] Hansen, B. (1996a). "Inference when a nuisance parameter is not identified under the null hypothesis," Econometrica, 64, 413-430.

[19] Hansen, B., (1996b) "Erratum" Journal of Applied Econometrics, 11, 195-198.

[20] Hansen B., (1999), "Sample Splitting and Threshold Estimation," Econometrica, forthcoming.

[21] Judge, G., Griffiths, W., Hill, R.C. and Lee, T.-C. (1985). The Theory and Practice of Econometrics, second edition, New York: John Wiley and Sons.

[22] Kim, C-J., and Nelson, C., (1999) State-Space Models with Regime-Switching:Classical and Gibbs-Sampling Approaches with Applications, MIT Press, Cambridge.

[23] Koop, G. (1996). "Parameter uncertainty and impulse response analysis," Journal of Econometrics, 72, 135-149.

[24] Koop, G. Pesaran, M.H. and Potter, S.M. (1996). "Impulse response analysis in nonlinear multivariate models," Journal of Econometrics, 74, 119-148.

[25] Koop, G. and Potter, S.M. (1999a). " Bayes factors and nonlinearity: evidence from economic time series," Journal of Econometrics, forthcoming.

[26] Koop, G. and Potter, S.M. (1999b). "Asymmetries in US Unemployment," Journal of Business and Economic Statistics, forthcoming.

[27] Lubrano, M. (1999). "Bayesian analysis of nonlinear time series models with a threshold" to appear in Nonlinear Econometric Modelling, William Barnett et al. (ed.), Cambridge: Cambridge University Press..

[28] Pesaran, M.H., and Potter, S. (1997). "A Floor and Ceiling Model of US Output," Journal of Economic Dynamics and Control 21, 661-695.

[29] Poirier, D. (1995). Intermediate Statistics and Econometrics, Cambridge: The MIT Press.

[30] Potter, S. (1995). “A Nonlinear approach to US GNP,” Journal of Applied Econometrics, 10, 109-125.

[31] Potter, S. (1999) “Nonlinear Impulse Response Functions,” Journal of Economic Dynamics and Control, forthcoming.

[32] Priestley, M.B., (1988), Non-Linear and Non-Stationary Time Series, New York: Academic Press

[33] Teräsvirta, T. and Anderson, H. (1992). ”Characterising nonlinearities in business cycles using smooth transition autoregressive models,” Journal of Applied Econometrics, S119-S136.

[34] Teräsvirta, T. (1994), “Specification, estimation and evaluation of smooth transition autoregressive models” Journal of the American Statistical Association, 89, 208-218.

[35] Tong, H. (1983) Threshold models in non-linear time series analysis, Lecture Notes in Statistics, No. 21, Springer, Heidelberg.

[36] Tong, H., (1990), Non-linear Time Series: A Dynamical System Approach, Oxford: Clarendon Press