EVALUATING THE EFFECTIVENESS OF STATE-SWITCHING TIME SERIES MODELS FOR U.S. REAL OUTPUT*

Richard A. Ashley
Department of Economics
Virginia Tech (VPI)

Douglas M. Patterson
Department of Finance
Virginia Tech (VPI)

August 28, 2005

Abstract

Two types of state-switching models for U.S. real output have been proposed: models that switch randomly between states (as in Hamilton (1989)) and models that switch states deterministically, as in the threshold autoregressive model of Potter (1995). These models have been justified primarily on how well they fit the sample data, yielding statistically significant estimates of the model coefficients. Here we propose a new approach to the evaluation of an estimated nonlinear time series model which provides a complement to existing methods based on in-sample fit or on out-of-sample forecasting. In this new approach, a battery of distinct nonlinearity tests is applied to the sample data, resulting in a set of p-values for rejecting the null hypothesis of a linear generating mechanism. This set of p-values is taken to be a "stylized fact" characterizing the nonlinear serial dependence in the generating mechanism of the time series. The effectiveness of an estimated nonlinear model for this time series is then evaluated in terms of the congruence between this stylized fact and a set of nonlinearity test results obtained from data simulated using the estimated model. In particular, we derive a portmanteau statistic based on this set of nonlinearity test p-values which allows us to test the proposition that a given model adequately captures the nonlinear serial dependence in the sample data. We apply the method to several estimated state-switching models of U.S. real output.

JEL classification: E32, C22, C40

*We are grateful to Anthony Atkinson, Ivan Pastine, and Douglas Steigerwald for helpful comments and to M.J. Hinich and B. LeBaron for sharing their computer codes with us. The authors may be contacted at ashleyr@vt.edu and amex@vt.edu, respectively.

1. Introduction

In the 1970s and 1980s a time series like U.S. real output would typically be modeled as linear function of its own past, using an ARMA or VAR framework. By the 1990's, however, it was widely recognized that nonlinear serial dependence is an essential feature in such time series. Stimulated by contributions from Tong (1983) and Hamilton (1989), this nonlinear serial dependence was (and remains) most commonly modeled using state-switching (or “regime-switching”) models.

In Tong’s threshold autoregression (TAR) framework, the time series switches deterministically from one linear autoregressive model to another based on the lagged value of an observed variable, with the parameters (including the threshold value at which switching occurs) estimated via nonlinear least squares. In Hamilton’s Markov switching framework, the time series switches from one linear autoregressive model to another at random, with the parameters (including the state transition probabilities) typically estimated via maximum likelihood methods. Each framework, of course, has been elaborated in various ways. In both frameworks, however, the adequacy of the model is primarily predicated on the statistical significance of the relevant coefficient estimates relative to their asymptotic standard error estimates – in essence, the model is accepted because it fits the sample data reasonably well and because more elaborate parameterizations do not materially improve the fit.

Here we use a new model validation approach to examine whether published implementations based on either of these frameworks is adequate to explain the observed nonlinear serial dependence in U.S. real output. The new approach is described in the next section. It takes the pattern with which a battery of distinct tests for nonlinear serial dependence reject the null hypothesis of linear serial dependence as a new "stylized fact" about the time series and examines the degree to which a particular estimated nonlinear model for the time series is capable of reproducing this pattern. The set of nonlinearity tests used here is described in Section 3. In Section 4 these tests are used to generate this new "stylized fact" about U.S. real output; four estimated models for U.S. real output are then described and subsequently analyzed using the new approach. Two of these estimated models (due to Lam (1997)) are Markov state-switching models; one assumes fixed state-transition probabilities (as in Hamilton (1989)) and the other assumes that the transition probabilities depend on how long the economy has been in that state. The other two estimated models are deterministic state-switching models. The first of these is a threshold autoregressive model for U.S. real output estimated by Potter (1995); the second of these is a "smooth transition autoregressive" or "STAR" variation on this model (as in Teräsvirta, T. and H. Anderson (1992)), in which the economy switches smoothly from one linear model to another based on the lagged value of real output.

The approach proposed here – in contrast to that of Harding and Pagan (2002), for example – yields a statistical test of the proposition that a particular model generated the observed sample data. Applying the new approach to the four models of U.S. real output noted above, this test indicates that the estimated STAR and Markov switching models are broadly consistent with the results obtained from applying the battery of nonlinearity tests to the sample data, but allows us to reject the threshold autoregressive specification at the 5% level.

2. A New Approach for Evaluating Nonlinear Time Series Models

Here we introduce a new approach for evaluating the effectiveness of an individual nonlinear model of a times series or for comparing the effectiveness of two such models. In this section we first discuss the need for this new methodology. Existing methods are then briefly reviewed, after which the new approach is described, along with some of its advantages as a complement to existing methods. This new technique is then applied in Sections 3 and 4 to evaluate the effectiveness of several state-switching models for U.S. real output.

a. Why evaluation of nonlinear time series models is important.

The proposition that efficient parameter estimation and valid statistical inference hinge crucially on appropriate model specification is hardly controversial. Further, ample theoretical and empirical evidence indicates that nonlinear generating mechanisms are important in a number of macroeconomic and financial processes.

For example, many theoretical macroeconomic models are highly nonlinear, from Hicks' (1950) elaboration of the Samuelson multiplier-accelerator theory, to Grandmont's (1985) overlapping generations model, to labor hoarding models such as Hall (1990), and to recent models, such as Palm and Pfann (1997), which are based on an explicit treatment of asymmetric adjustment costs. The nonlinearity in these models is intrinsic to the macroeconomic hypotheses embodied in them and is essential to the derivation of the key properties these models display, such as asymmetric business cycles and chaotic dynamics – see Barnett and Hinich (1992) and Barnett, *et al.* (1995).

Moreover, numerous empirical studies have found statistical evidence for significant nonlinearity in the generating mechanisms of important macroeconomic and/or financial time series. Examples include Engle (1982), Tong (1983), Hinich and Patterson (1985), Tsay (1986), Ashley and Patterson (1989), Hamilton (1989), Brock, Hsieh and LeBaron (1991), Potter (1995), and Altug, Ashley and Patterson (1999) among many others. Some of these studies have simply detected nonlinearity in a particular time series – e.g., Ashley and Patterson (1989). Other studies – e.g., Hamilton (1989) and Potter (1995) – assume at the outset that the nonlinearity takes a particular form and estimate an explicit model for the nonlinear serial dependence.

Despite several attempts – e.g., Priestley (1988) and Gallant and Nychka (1987) – the field of nonlinear time series analysis lacks a widely accepted model identification algorithm analogous to that proposed by Box and Jenkins (1976) for linear processes. Consequently, it is entirely possible for each of several research groups to start with the same time series data and produce substantially different models to represent its true generating mechanism, simply because each group begins from a different assumption as to the family of nonlinear processes that generated the data.

For example, one group might assume that the true generating mechanism is a threshold autoregression, whereas the other group might assume a Markov switching mechanism. The resulting fitted models will be quite different from one another, so they can't both be correct.

In fact, one or neither specification might be reasonably close to the correct data generating mechanism. How can one objectively assess the relative and absolute value of these models as approximations to the true generating mechanism?

b. Existing approaches for evaluating time series models.

A common approach is to ask which model fits the data best, based on R_c^2, FPE, AIC, BIC, etc. Sample fit is important, but since the sample data are customarily (and necessarily) mined to identify the particular form of whatever kind of model is being considered, the fact that the resulting model fits the data well usually reflects the flexibility of the framework being used (threshold autoregressive, Markov switching, neural net, or whatever) more than it does which kind of model is closer to the specification which actually generated the data.

Another approach relies on relative out-of-sample forecasting effectiveness as a criterion for model choice. Out-of-sample forecasting can give substantially credible support to a particular model or to one model specification over another. But the results from this approach can be idiosyncratic to the particular model validation period chosen unless the hold-out sample is lengthy, in which case an insufficient number of observations may remain for model specification and estimation. (In particular, on might expect that an adequate postsample forecast period for evaluating a state switching model would need to be sufficiently long as to include a number of state switches.) Quite often, moreover, one finds that neither of two candidate nonlinear time series models provides out-of-sample forecasts which are very useful; in such cases it seems unreasonable to prefer one model over the other on this basis. Such poor out-of-sample forecasting can arise because both model specifications are totally inadequate, but it can also reflect the fact that forecasts from nonlinear models are very sensitive to even modest model mis-specification. In other words, it might be the case that one model is far closer to the true data generating mechanism in the ways we most care about, yet neither model is close enough to forecast out-of-sample creditably well.

c. A new approach

Here we introduce a new approach – complementary to the “sample fit” and “out-of-sample forecasting” approaches outlined above – for either evaluating an individual nonlinear times series model or comparing two such models. Our approach is based on a battery of distinct nonlinearity tests.

We briefly discuss a selection of nonlinearity tests in Section 3; more complete descriptions are given in Appendix 1. The reason that there are so many tests (and the reason that no comprehensive model identification algorithm for nonlinear models has found widespread acceptance) is that there are many distinctly different ways in which the current value of a time series can depend nonlinearly on its own past. Consequently, many tests for nonlinearity can be constructed, each focusing on a different aspect or effect of nonlinear serial dependence – e.g, one test might focus on the way nonlinear serial dependence impacts the higher order moments of the time series, whereas another test might look at how close different sequential m-tuples of the process are to each other. Thus, some nonlinearity tests will naturally be substantially more powerful than others against specific alternatives.

Our approach leverages this diversity by taking the pattern of p-values with which a set of nonlinearity tests rejects the null hypothesis of a linear generating mechanism for a particular times series as a new stylized fact characterizing the nonlinear serial dependence in this time series. One can then ask of any estimated model for this time series, “How well does it reproduce this stylized fact?”

Thus, our approach is similar in spirit to the more descriptive examination by Harding and Pagan (2002) of how well a statistical model is able to track specific features of the shape of the business cycle. Indeed, if one includes explicitly shape-related tests – e.g., the tests for steepness and depth proposed by Ramsey and Rothman (1996), Verbrugge (1997), and others – in the set of nonlinearity tests considered, then our approach subsumes and extends theirs.

One could simulate data from the estimated model and compute the power of each nonlinearity test to reject the null hypothesis of a linear generating mechanism against this particular alternative generating mechanism. If the estimated model is effective at modeling the nonlinear serial dependence in the actual data, then one would expect that the tests which are most powerful in detecting this alternative are the ones which reject the null hypothesis with the lowest p-values using these data. In contrast, if the tests which provide the strongest evidence for nonlinearity are ones with relatively small power to detect the kind of nonlinearities generated by this model, it seems less plausible that the actual generating mechanism for these data is of this kind.

Our approach takes this reasoning one step further, allowing us to construct a statistical test of the proposition that a specific nonlinear model is capturing the nonlinear serial dependence in the data, as distinct from merely fitting the sample data well in a least squares sense.

Suppose that *r* nonlinearity tests have been applied to the sample data, yielding *r* p-values $\left(p_1^{obs} \ldots p_r^{obs}\right)$ for rejection of the null hypothesis of a linear generating mechanism for the time series. Consider, then, a "portmanteau" test statistic quantifying the discrepancy between this set of results and the set of p-values one might expect had the sample data been generated by this specific model:

$$AP\left(p_1 \ldots p_r\right) = \sum_{i=1}^{r} \left[p_i - E\{p_i\}\right]^2$$

Note that the expectation in this expression is taken over the joint distribution of the *r*-vector $\left(\mathrm{p}_1 \ldots \mathrm{p}_\mathrm{r}\right)$. This vector is a random variable because the p-value for each of the *r* nonlinearity tests is a monotonic transformation of the test statistic for that particular nonlinearity test. Thus, for example, one might expect the inverse error function of a p-value from the BDS test to be asymptotically a unit normal under the (counterfactual) supposition that the data were serially independent.

Both this expectation and the sampling distribution of the AP test statistic are readily obtained by monte carlo simulation under the null hypothesis that the sample data are generated by any particular model – indeed, these simulations are already done in calculating the power of the individual nonlinearity tests for this model. The p-value at which one can reject this null hypothesis is thus just the fraction of these simulations which yields AP values in excess of $\mathrm{AP}\left(\mathrm{p}_1^{\mathrm{obs}} \ldots \mathrm{p}_\mathrm{r}^{\mathrm{obs}}\right)$.

Note that the *r* nonlinearity tests used in the AP test statistic defined above need not be in any sense optimal. In fact, the result of a given nonlinearity test is potentially an informative feature of a “stylized fact” characterizing a particular time series whenever the power of this nonlinearity test against the particular generating mechanism being considered differs substantially from this test’s power to detect the kind of nonlinear serial dependence actually present in the time series. In particular, even a nonlinearity test with low power to detect the nonlinear serial dependence actually present in the sample data can be very informative in our framework. For example, suppose that a threshold autoregressive (TAR) model has been proposed and estimated for a particular time series and that it is found that this nonlinearity test has very high power to detect the nonlinear serial dependence generated in data simulated from this particular TAR model. If – as is likely – this nonlinearity test fails to reject its null hypothesis using the sample data, then it would be generating potentially substantial evidence against the proposition that the proposed TAR model is an adequate approximation to the actual generating mechanism for this time series.

Note also that this testing procedure does not depend on a detailed knowledge of how the estimated model was obtained. Our method is therefore applicable to models which have been estimated by intricate (or even proprietary) methods, so long as one can either simulate the estimated model oneself or obtain a long realization from someone who can and has.

Thus, the results from this new approach

- are not systematically distorted by the relative or absolute amount of specification search (data mining) available or utilized in the production of either model, since the approach is based on the estimated model's ability to replicate the nonlinearity properties observed in the sample data rather than on its ability to fit the sample data in a least squares sense,
- do not require specification of a "hold-out" sample for model validation and do not hinge on the ability of either model to successfully forecast outside of its specification/estimation period,

and

- do not require re-estimation of the model – all that is required is a sufficiently long simulated realization.

d. Plan of the rest of the paper.

The particular selection of nonlinearity tests used in this paper is briefly described in Section 3; there we present power estimates indicating that these tests are distinct from one another in the sense that at least some of them have differing relative power to detect commonly considered nonlinear processes. In Section 4 the new approach described in this Section is applied to evaluate the effectiveness of several state-switching models for U.S. real output.

3. A Selection of Nonlinearity Tests

The six nonlinearity tests used below are listed and briefly described in Table 1 and below; since these tests are well known, more complete descriptions are given in Appendix 1.

Table 1
Nonlinearity Tests Considered

Test	Focus	Reference
McLeod and Li	ARCH/GARCH	McLeod and Li (1983)
Engle LM	ARCH/GARCH	Engle (1982)
BDS	General serial dependence	Brock, Dechert, and Scheinkman (1996)
Tsay	Quadratic terms (time domain)	Tsay (1986)
Hinich Bicovariance	3^d order moments (time domain)	Hinich and Patterson (1995) and Hinich (1996)
Hinich Bispectrum	3^d order moments (frequency domain)	Hinich (1982)

With the exception of the Hinich Bispectrum test, each of these procedures is actually testing for serial dependence of any kind, whether linear or nonlinear. Consequently, data pre-whitening is necessary prior to the application of each of these tests, in order to eliminate any linear serial dependence in the data. Since each of these tests is only asymptotically justified, bootstrapping is in each case also necessary in order to obtain a correctly sized test. These issues have been dealt with elsewhere – e.g. in Patterson and Ashley (2000) – and hence are summarized here in Appendix 2, along with updated simulations confirming that these tests as implemented here are correctly sized for the sample lengths used in the estimated models for real output analyzed in Section 4 below. It bears mention, however, that – even with bootstrapping – the BDS test is correctly sized only for embedding dimension (m) equal to two. The size of the BDS test at higher values of m is apparently distorted by a high sensitivity to the minor amounts of linear dependence remaining in the series on those occasions where the pre-whitening procedure mis-identifies the order of the AR(p) process used to eliminate serial correlation in the series prior to bootstrapping. Consequently, all BDS test results quoted below are for m equal to two.

Many other tests for nonlinear serial dependence have been described in the literature, including: Ramsey (1969), Ashley and Patterson (1986), Saikkonen and Luukkonen (1988), White (1989), Mizrach (1991), Nychka, et al. (1992), Kaplan (1993), Dalle Molle and Hinich (1995), and Hansen (1999). Since asymmetry is a common consequence of nonlinear serial dependence, one might also consider tests for steepness or deepness, as in Ramsey and Rothman (1996) and Verbrugge (1997). No representation is made here – nor, for the present purpose, needs to be made – that the group of tests listed in Table 1 is in any sense optimal nor that these tests in any well-defined sense "span the space" of all possible nonlinearity tests. Indeed, insofar as useful new tests for nonlinear serial dependence continue to appear and insofar as some usefully distinct existing tests have no doubt been omitted from consideration here, our results using this group of tests can be taken as a lower limit on the potential usefulness of the proposed approach.

Rather, the issue here is the degree to which each test in the group has power to detect some distinct aspect of nonlinear serial dependence. This issue is briefly examined in this section by estimating the power of each test in the group against a number of alternative data generating processes commonly considered in the literature. Other studies examining the ability of various tests to detect nonlinearity include: Lee, et al. (1993), Barnett, et al. (1997), and Lemos and Stokes (1998).

The processes considered here are listed in Table 2. In each case, the innovation series (ϵ_t) is an independent unit normal variate; Student's t, exponential, and symmetric Paretian variates are considered in Patterson and Ashley (2000), yielding similar results. The nonlinear autoregressive model used here is taken from Lee et al. (1993). As is the case with the set of nonlinearity tests considered, no representation is made here that the set of processes included in Table 2 in any well-defined sense encompasses all possible nonlinear generating mechanisms. On the other hand, the set of processes given in Table 2 does include generic versions of a number of different nonlinear models which have received empirical and/or theoretical attention in the literature.

Table 2. Data Generating Processes Considered

Conditional Heteroskedasticity Models:	
ARCH	$x_t = (h_t)^{.5} \epsilon_t$ $h_t = .000019 + .846\{x^2_{t-1} + .3x^2_{t-2} + .2x^2_{t-3} + .1x^2_{t-4}\}$
GARCH	$x_t = (h_t)^{.5} \epsilon_t$ $h_t = .011 + .12\,(x_{t-1})^2 + .85\,h_{t-1}$
Switching Models:	
Threshold Autoregression (TAR)	$x_t = -.5\,x_{t-1} + \epsilon_t$ if $x_{t-1} < 1$ $x_t = .4\,x_{t-1} + \epsilon_t$ otherwise
Two State Markov Switching	$x_t = -.5\,x_{t-1} + \epsilon_t$ if in state 1 $x_t = .4\,x_{t-1} + \epsilon_t$ if in state 2 (Remain in state with probability .90)
Other models:	
Bilinear	$x_t = .7\,x_{t-1}\,\epsilon_{t-2} + \epsilon_t$
Nonlinear Autoregressive	$x_t = .7\,\lvert x_{t-1}\rvert \,/\, (.7\,\lvert x_{t-1}\rvert + 2) + \epsilon_t$

The estimated power of each nonlinearity test against each of these alternatives is given in Table 3 below. All figures quoted are based on 250 generated samples; the parameters L, p, m, ℓ, k, and M are defined in Appendix 1, where each test is discussed. BDS test results were calculated for ϵ equal to one half, one, and two standard deviations; for brevity, results are quoted only for ϵ equal to one; BDS test results at values of the embedding dimension (m) exceeding two are omitted due to the problems with the size of the test at these embedding dimensions noted earlier in this section. The 5% critical region for each test was obtained using 1000 bootstrap replications; details on the pre-whitening and bootstrapping procedures used are given in Appendix 2.

Table 3 Power Estimates of 5% Tests
200 Observations

	McLeod-Li	Engle LM	BDS	Tsay	Bicovariance	Bispectrum
	L = 24	p = 5	m = 2	k = 5	ℓ = 8	M = 24
Conditional Heteroskedasticity Models:						
ARCH	.62	.80	1.00	.32	.46	.10
GARCH	.71	.72	.65	.38	.68	.09
Switching Models:						
Threshold AR	.12	.13	.62	.78	.10	.12
Markov	.17	.32	.56	.11	.13	.06
Other Models:						
Bilinear	.85	.98	.97	.99	.99	.15
Nonlinear AR	.04	.06	.06	.09	.09	.06

The power results in Table 3 indicate that no single test dominates the others across all six alternative generating processes. For present purposes, however, what is important is that this set of tests displays distinct patterns of power against these specific alternatives. For example, both the BDS and Tsay tests seem to be notably powerful against the threshold autoregressive alternative, whereas the BDS test appears to be uniquely powerful against this Markov switching alternative.

Of course, the most relevant set of nonlinear processes to consider are the ones which have actually been specified and estimated in the literature for the time series at issue: U.S. real output, in the present case. Several such processes are analyzed in the next section.

4. Evaluating Four State-Switching Models for U.S. Real GNP

In this section we apply our new approach to analyze the effectiveness of four state-switching models for the quarterly growth rate of U.S. real GNP. One of these is a threshold autoregressive model due to Potter (1995). Another two are Markov switching models due to Lam (1997); one of these Markov switching models assumes constant state transition probabilities {as in Hamilton (1989), only with a longer sample period} and the other models each state transition probability as a function of the number of periods that the system has been in the current state. In addition, we have estimated our own Smooth Transition Autoregressive (L-STAR) model for U.S. real GNP.

The battery of tests discussed in Section 3 were applied to the logarithmic growth rate of U.S. real GNP over a sample period similar to that used by Potter and Lam in specifying and estimating their models for this time series. The resulting p-values for rejection of the null hypothesis of a linear generating mechanism for this series are given in Table 4:

Table 4 Significance Levels for Nonlinearity Tests on U.S. Real GNP {1953I to 1993III}

McLeod-Li	Engle LM	BDS	Tsay	Bicovariance	Bispectrum
.218	.525	.356	.025	.017	.331

Consistent with results in Ashley and Patterson (1989) on the U.S. Index of Industrial Production and with results in Hamilton (1989), Potter (1995), and Altug, et al. (1999) on real GNP itself, the null hypothesis of a linear generating mechanism for this time series can be rejected at the 5% level. The *pattern* of these test results – which we here take to be a new stylized fact about U.S. real GNP – is noteworthy, however.

GNP is used here rather than GDP for consistency with the estimated models to be analyzed; these two series in any case differ little for U.S. data. More problematic is the fact that both GNP and GDP figures have been repeatedly revised since the time these models were estimated; even the methodology for deflating GNP has changed subsequent to Potter's work. Fortunately, our method does not require us to re-estimate these models, but it does seem important to compute the nonlinearity test p-values using consistently revised GNP data similar to that which was available to Potter and Lam at the time. Consequently, the data used in Table 4 is drawn from a sample obtained in early 1994 for use in Altug, et al. (1999). We note in passing that the most striking aspect of the results in Table 4 – that the p-values for the Tsay and Hinich

Bicovariance are notably smaller than that for the BDS test – is still evident, albeit in somewhat muted form, using current chain-weighted real GDP figures over this sample period: the p-values for these three tests are in that case .090, .067, and .012, respectively.

McConnell and Perez-Quiros (2000) observe that the variance in U.S. real output dropped in the early to middle 1980's, but this shift is by no means evident in the data over the sample periods used by Potter and Lam. Nor is the pattern of nonlinearity test p-values observed in the actual real output data an artifact of this shift: using sample data from 1953I to 1983IV yields similar results, although all of the p-values are much larger because ending the sample period in 1983 reduces the sample length from 163 to 124 observations. In particular, the BDS test still fails to reject the null hypothesis of linearity (p-value = .35), whereas the Tsay and Hinich Bicovariance tests, with p-values of .13 and .08, respectively, still yield some evidence against the null hypothesis of a linear generating mechanism.

Real output is commonly modeled as some sort of two-state regime switching process nowadays. Note, however, that the pattern of significance levels (p-values) in Table 4 is quite different from what one might expect based on the regime switching model power results reported in Table 3. For example, using data generated from the simple TAR model considered in Section 3, the Hinich Bicovariance test has quite low power and the BDS test has relatively high power. Similarly, for data generated from the simple Markov regime switching model considered in Section 3, both the Tsay and the Hinich Bicovariance tests have low power relative to the BDS test. In contrast, using the actual data on real GNP, we see in Table 4 a fairly strong rejection from the Tsay test and the Hinich Bicovariance test, but the BDS test cannot reject the null hypothesis at all. Thus, if the true generating mechanism for real GNP is a regime-switching model similar to either model considered in Section 3, it is surprising that the p-value for the

BDS test using the sample data is so high relative to the p-values for the Tsay and Hinich Bicovariance tests.

Of course, what is really at issue is not whether the regime-switching models examined in the power calculations of Section 3 are consistent with this new stylized fact about U.S. real GNP, but whether or not regime-switching models estimated using actual U.S. real GNP data are consistent with it.

To examine this issue, we first estimate the power of all six tests using simulated data generated from each of the four estimated models for real GNP, all of which were specified and estimated over essentially the same sample period used in obtaining the results given in Table 4. If any one of these estimated models is consistent with our new "stylized fact" concerning real GNP, then the tests which did reject the null hypothesis of a linear generating mechanism for the actual data (i.e., the Tsay and Hinich Bicovariance tests) should have relatively high power to detect this particular alternative, and the tests (i.e., BDS) which failed to reject the null hypothesis of a linear generating mechanism in the actual data should have relatively low power to detect this particular alternative. For each estimated model we then obtain a statistical test of the null hypothesis that this model generated the sample data. This null hypothesis is tested by computing the percentage of artificial data sets simulated from the estimated model which yield a set of nonlinearity test results which differ from those expected from this generating model to a greater degree than does the set of test results (given in Table 4) observed using the actual sample data.

The first model considered is a threshold autoregressive model for U.S. real GNP identified and estimated by Potter (1995), based on an identification procedure suggested by Tsay (1991). Potter's preferred model is:

$$y_t = \underset{(.423)}{-.808} + \underset{(.185)}{.516}\, y_{t-1} - \underset{(.353)}{.946}\, y_{t-2} + \underset{(.216)}{.352}\, y_{t-5} + \epsilon_t \qquad \text{for } y_{t-2} \leq 0$$

$$y_t = \underset{(.161)}{.517} + \underset{(.080)}{.299}\, y_{t-1} + \underset{(.107)}{.189}\, y_{t-2} - \underset{(.069)}{1.143}\, y_{t-5} + \eta_t \qquad \text{for } y_{t-2} > 0$$

where the figures in parentheses are estimated standard errors and the sample period is 1948III to 1990IV. Potter's sample period includes a number of large variations at the beginning of the sample. In common with most authors, we exclude these unusual observations by starting our sample period for Table 4 somewhat later. Re-estimating Potter's model over the period 1953I to 1993III, however, yields similar results, except that the term at lag five is less significant.

The second model considered here is a logistic Smooth Transition Autoregressive (L-STAR) model, as described in Teräsvirta, T. and H. Anderson (1992), Granger and Teräsvirta (1993), and Teräsvirta, T. (1994). Using the sample period 1953I to 1993III and starting with Potter's TAR model specification, we obtain

$$y_t = \underset{(.131)}{.226} + F(y_{t-2})\,[\underset{(.154)}{.158}\, y_{t-1} - \underset{(.183)}{.363}\, y_{t-2}]$$

$$+ \; [1 - F(y_{t-2})]\,[\underset{(.090)}{.315}\, y_{t-1} - \underset{(.115)}{.262}\, y_{t-2}] + e_t$$

where F(x) is $1/(1 + e^{.589x})$. The terms at lag five in Potter's specification are omitted because they are not statistically significant and yield models which are inferior using either the AIC or BIC criterion.

The third model considered here is a two-state Markov switching model first proposed by Hamilton (1989) and re-estimated by Lam (1997) over the sample period 1952II to 1996IV:

$$y_t = \left\{ \underset{(.093)}{.852}\ (\text{State I})\ \text{or}\ \underset{(.453)}{-1.500}\ (\text{State II}) \right\}$$

$$+ \underset{(.084)}{.388}\, y_{t-1} + \underset{(.102)}{.097}\, y_{t-2} - \underset{(.099)}{.106}\, y_{t-3} - \underset{(.083)}{.127}\, y_{t-4} + \epsilon_t$$

where the system remains in State I with probability .966 and remains in State II with probability .208.

Finally, the fourth model considered here is a two-state Markov switching model proposed by Lam (1997) which generalizes this framework to allow both the mean growth rate and the transition probabilities to depend on D_t, the number of quarters the system has been in its current state:

$$y_t = \underset{(.1198)}{1.6250} - \underset{(.0241)}{.0775}\,(D_t - 1) + \underset{(.0006)}{.0013}\,(D_t - 1)^2 \qquad (\text{State I})$$

$$\underset{(.1574)}{-.2755} - \underset{(.2237)}{1.2302}\,(D_t - 1) + \underset{(.0884)}{.6267}\,(D_t - 1)^2 \qquad (\text{State II})$$

$$+ \underset{(.099)}{.385}\, y_{t-1} + \underset{(.134)}{.401}\, y_{t-2} - \underset{(.099)}{.292}\, y_{t-3} - \underset{(.114)}{.193}\, y_{t-4} + \epsilon_t$$

where the probability of remaining in State I is the logistic of

$$\underset{(.3861)}{.9867} + \underset{(.0342)}{.0970}\,(D_t - 1)$$

and the probability of remaining in State II is the logistic of

$$\begin{array}{ccc} 2.0873 & - & 1.6992\ (D_t - 1). \\ (.7660) & & (.5861) \end{array}$$

Table 5 lists power estimates for the six nonlinearity tests obtained using 1000 artificial samples generated using each of these estimated models; the sample length for each generated series was chosen to match the number of observations in the actual sample data used to obtain

Table 5. Empirical Power of 5% Tests Using Data
Generated From Estimated Models for U.S. Real GNP

McLeod-Li	Engle LM	BDS	Tsay	Bicovariance	Bispectrum
Using simulated data from Potter (1995) TAR model for U.S. real GNP:					
.91	.93	.83	.95	.96	.02
Using simulated data from estimated L-STAR model for U.S. real GNP:					
.04	.06	.06	.13	.05	.09
Using simulated data from Lam (1997) re-estimation of Hamilton (constant transition probabilities) Markov switching model for U.S. real GNP:					
.05	.07	.11	.07	.07	.07
Using simulated data from Lam (1997) estimated Markov switching model for U.S. real GNP with duration dependent mean and transition probabilities:					
.09	.11	.11	.13	.14	.06

the test results given in Table 4. In each case the model innovations used were independent draws from a normal distribution with variance equal to the estimated model error variance; this corresponds to what is usually known as the parametric bootstrap.

Except for the Hinich Bispectrum test, all of the tests appear to have high power to detect the nonlinearity in artificial data generated from Potter's estimated TAR model. With power this high, one would expect the McLeod-Li, Engle LM, and BDS tests to reject the null hypothesis of linearity in the actual data if it were generated by a model similar to Potter's TAR, but reference to Table 4 shows that these three tests do not reject this null hypothesis in the sample data. In contrast, the power results in Table 5 indicate that none of the tests seems particularly effective at detecting the nonlinearity in artificial data generated using either the L-STAR or one of the two Markov switching models estimated by Lam (1997). With power this low, one would not expect the Tsay and Hinich Bicovariance tests to reject linearity in the actual data if these data were generated by an L-STAR or by a Markov switching model such as these; but reference to Table 4 shows that they do reject this null hypothesis. Thus, the pattern of which tests reject linearity using the actual real GNP data conflicts with the pattern of power results obtained using artificial data generated from each of the four estimated models.

In principle, these discrepancies could be due to ordinary sampling variation. To assess the statistical significance of these discrepancies for a particular estimated generating model, we applied the test described in Section 2, approximating the joint distribution of the six nonlinearity test p-values (p_1 ... p_6) by the set of 1000 p-value vectors obtained using the 1000 artificial data sets generated from this model. We then computed the fraction of these 1000 p-value vectors yielding a larger AP test statistic than that yielded by the p-value vector (quoted in Table 4) obtained using the sample data.

Table 6 summarizes the results of these calculations using several variations of the AP test statistic. The first row of numbers in this table gives the results obtained using the AP test statistic exactly as defined in Section 2. The remaining rows display analogous alternative results obtained using the sum of squared deviations from the median p-value for each of the six nonlinearity tests instead of the mean values of these p-values and from using the sum of the absolute deviations instead of sum of the squared deviations.

Table 6 P-values for Rejecting Models of US GNP Using AP Test

	Markov Switching Models		Deterministic Switching Models	
	Constant transition probabilites	Duration-dependent transition probabilities	Threshold Autoregression Model	Smooth Transition Threshold Autoregression Model
Squared Deviation AP Test Statistic				
deviation from mean	.425	.648	.058	.513
deviation from median	.472	.735	.053	.560
Absolute Deviation AP Test Statistic				
deviation from mean	.570	.644	.040	.575
deviation from median	.616	.719	.035	.631

In view of the modest sensitivity of these results to the form of the AP statistic, we also investigated an alternative formulation in which each of the six test p-values used is transformed by the inverse logit cumulative distribution function:

$$\tilde{AP}\left(p_1 \dots p_6\right) = \sum_{i=1}^{6} \left[g\left(p_i\right) - E\left\{g\left(p_i\right)\right\}\right]^2$$

where g(p) is ln{p/(1-p)}. Again, the test was also computed in terms of absolute rather than squared deviations and in terms of deviations from the median rather than deviations from the mean:

Table 7 P-values for Rejecting Models of US GNP Using AP Test Applied to Logit-transformed Nonlinearity Test p-values

	Markov Switching Models		Deterministic Switching Models	
	Constant transition probabilites	Duration-dependent transition probabilities	Threshold Autoregression Model	Smooth Transition Threshold Autoregression Model
Squared Deviation $\tilde{AP}$ Test Statistic				
deviation from mean	.211	.343	.035	.271
deviation from median	.208	.315	.033	.244
Absolute Deviation $\tilde{AP}$ Test Statistic				
deviation from mean	.337	.415	.033	.357
deviation from median	.333	.402	.031	.346

All four variations on both versions of the test yield essentially the same results. However, the $\tilde{AP}$ results based on the logit-transformed nonlinearity test p-values are a bit stronger and are notably more stable across the four variations; consequently, these results are preferable.

The $\tilde{AP}$ test fails to reject the STAR model and either of the Markov switching models. In contrast, all four variations of the $\tilde{AP}$ test statistic yield results indicating that the nonlinear serial dependence generated by the Potter threshold autoregressive model is inconsistent – at the 5% level of significance – with the pattern of nonlinearity test results obtained using the sample data. This is a fairly strong result in view of the fact that the 163 observations used in this analysis would be considered a small sample for an analysis of nonlinear serial dependence.

In view of the fact (Table 5) that all six nonlinearity tests have low power to detect the kind of nonlinear serial dependence induced by the L-STAR and Markov switching models, whereas the results in Table 4 show that the Tsay and Bicovariance tests do reject this null hypothesis at the 5% level using the sample data, it is somewhat surprising that neither the AP nor the $\tilde{AP}$ test is able to reject these models. The power of the AP test and $\tilde{AP}$ tests is apparently smaller in this situation.

5. Summary and Conclusions

As noted in Section 2, time series models are currently evaluated based on two criteria:

(1) the goodness of the model's fit to the sample data (often narrowed to a consideration of the statistical significance of the parameter estimates),

and

(2) the model's postsample forecasting ability.

Existing nonlinear modeling frameworks (TAR, STAR, Markov-switching, etc.) are sufficiently flexible as to routinely produce models with significant parameter estimates, but these approaches ordinarily produce models which fail to forecast postsample. (Perhaps reasonably so, since postsample forecasting periods are usually not sufficiently long as to include a substantial number of state switches.) Here we have here proposed a complementary evaluation strategy: we examine simulated data from an estimated model for the time series to see how well it reproduces the nonlinear serial dependence observed in the sample data. Because this observed nonlinear dependence is usually detected by means of nonlinearity tests based on higher moments of the model errors, our approach is similar in spirit to the common practice of examining the sample correlogram of the errors (and the squares of the errors) made by a linear model.

Harding and Pagan (2002) and Morley and Piger (2004) have informally examined the degree to which simulated data from models for U.S. GDP yield business cycles with characteristics – e.g., asymmetry, duration, etc. – similar to those observed in the sample data. Our approach subsumes theirs in that one could include nonlinearity tests – e.g., asymmetry tests, such as those proposed by Ramsey and Rothman (1996) and Verbrugge (1997) – sensitive to these characteristics in the battery of nonlinearity tests used. In addition, our framework produces a straightforward statistical test of the model specification, whereas theirs does not.

In particular, this paper has employed our new approach to evaluate the ability of four estimated state switching models for U.S. real output – two Markov switching models due to Lam (1997), a threshold autoregressive (TAR) model due to Potter (1995), and a smooth transition autoregressive (STAR) model estimated here – to capture or explain the nonlinear serial dependence observed in this time series over the sample period 1953I to 1993III. The Markov switching and STAR models are not rejected, but we are able to reject the Potter threshold autoregressive model at the 5% level of significance.

As with the analysis of a sample correlogram of model fitting errors in a linear ARMA modeling setting, our model evaluation procedure is not intended to be an ending point. Rather, it is intended to form a diagnostic part of an iterative model specification process. In this case, our results suggest that either the specification of this particular TAR process needs modification (e.g., different lag structures, a third state, etc.) or that a smoother switching process (Markov switching, bilinear, or STAR) might be more suitable for modeling the nonlinear serial dependence in U.S. real output over this sample period.

References

Altug, S., Ashley, R., and Patterson, D. M. (1999). "Are Technology Shocks Nonlinear?" *Macroeconomic Dynamics 3(4)*, 506-533.

Ashley, R. and Patterson, D. M. (1986). "A Non-Parametric, Distribution-Free Test For Serial Independence In Stock Returns" *Journal of Financial and Quantitative Analysis 21*, 221-227.

Ashley, R. and Patterson, D. M. (1989). "Linear Versus Nonlinear Macroeconomies" *International Economic Review 30*, 685-704.

Ashley, R., Patterson, D. M. and Hinich, M. (1986). "A Diagnostic Test for Nonlinear Serial Dependence in Time Series Fitting Errors" *Journal of Time Series Analysis 7*, 165-78.

Barnett, W. A. and M.J. Hinich (1992) "Empirical Chaotic Dynamics in Economics," *Annals of Operations Research 37*, 1-15.

Barnett, W. A., A.R. Gallant, M.J. Hinich, J.A. Jungeilges, D.T. Kaplan, and M.J. Jensen (1995) "Robustness of Nonlinearity and Chaos Tests to Measurement Error, Inference Method, and Sample Size," *Journal of Economic Behavior and Organization 27*, 301-320.

Barnett, W. A., A.R. Gallant, M.J. Hinich, J.A. Jungeilges, D.T. Kaplan, and M.J. Jensen (1997) "A Single-Blind Controlled Competition Among Tests for Nonlinearity and Chaos" *Journal of Econometrics 82*, 157-92.

Bollerslev, Tim (1986) "Generalized Autoregressive Conditional Heteroskedasticity" *Journal of Econometrics 31,* 307-27.

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis* Holden-Day: San Francisco.

Brillinger, D. and M. Rosenblatt (1967) "Asymptotic Theory of kth Order Spectra" in *Spectral Analysis of Time Series*, (B. Harris, ed.) Wiley: New York, pp. 153-88.

Brock, W. A., Hsieh, D. A., and LeBaron, B.D. (1991) *A Test of Nonlinear Dynamics, Chaos, and Instability: Theory and Evidence* MIT Press: Cambridge.

Brock, W. A., Dechert W., and Scheinkman J. (1996) "A Test for Independence Based on the Correlation Dimension" *Econometric Reviews 15*, 197-235.

Dalle Molle, J. W. and Hinich M.J. (1995) "Trispectral Analysis of Stationary Random Time Series." *Journal of the Acoustical Society of America 97*, 2963-2978.

David, H. A. (1970) *Order Statistics* Wiley: New York.

Engle, Robert F. (1982) “Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation” *Econometrica 50,* 987-1007.

Fama, E. F. (1965) “The Behavior of Stock Market Prices” *Journal of Business 38*, 34-105.

Gallant, A. R. and D. W. Nychka (1987) “Seminonparametric Maximum Likelihood Estimation,” *Econometrica 55*, 363-90.

Grandmont, J. M. (1985) “On Endogenous Competitive Business Cycles” *Econometrica 53*, 995-1045.

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

Granger, C. W. J. and Teräsvirta, T. (1993) *Modelling Nonlinear Economic Relationships* Oxford University Press: Oxford.

Hall, R. (1990) “Invariance Properties of Solow’s Productivity Residual” in P. Diamond (ed.) *Growth/Productivity/Employment* MIT Press: Cambridge.

Hamilton, James (1989) “A New Approach to the Economic Analysis of Non-Stationary Time Series and the Business Cycle” *Econometrica 57*, 357-84.

Hansen, B. E. (1999) “Testing for Linearity” *Journal of Economic Surveys 13*, 551-576.

Harding, D. and Pagan, A. (2002) “Dissecting the Cycle: a Methodological Investigation” *Journal of Monetary Economics 49*, 365-81.

Hicks, J. R. (1950) *A Contribution to the Theory of the Trade Cycle* Oxford University Press: Oxford.

Hinich, M. (1982) “Testing for Gaussianity and Linearity of a Stationary Time Series” *Journal of Time Series Analysis 3*, 169-76.

Hinich, M. (1996) “Testing for Dependence in the Input to a Linear Time Series Model” *Journal of Nonparametric Statistics 6*, 205-221.

Hinich, M. and Patterson D. M. (1985) “Evidence of Nonlinearity in Daily Stock Returns” *Journal of Business and Economic Statistics 3*, 69-77.

Hinich, M. and Patterson D. M. (1995) “Detecting Epochs of Transient Dependence in White Noise,” unpublished manuscript, University of Texas at Austin.

Kanter, M. and Steiger W. L. (1974) “Regression and Autoregression with Infinite Variance” *Advances in Applied Probability 6*, 768-83.

Kaplan, D. T. (1993) “Exceptional Events as evidence for Determinism.” *Physica D 73*, 38-48.

Keenan, D.M. (1985) “A Tukey Nonadditivity-type Test for Time Series Nonlinearity.” *Biometrika 72*, 39-44.

Lam, P. (1997) “A Markov Switching Model of GNP Growth With Duration Dependence” (unpublished manuscript).

Lee, T., H. White, C.W.J. Granger (1993) “Testing for Neglected Nonlinearity in Time Series Models.” *Journal of Econometrics 56*, 269-90.

Lemos, M. and H. H. Stokes (1998) “A Single-Blind Controlled Competition Among Tests for Nonlinearity and Chaos; Further Results” unpublished manuscript, University of Illinois at Chicago.

de Lima, P. J. F. (1997) “On the Robustness of Nonlinearity Tests to Moment Condition Failure” *Journal of Econometrics 76*, 251-80.

McConnell, M. and G. Perez-Quiros (2000) “Output Fluctuations in the United States: What Has Changed Since the Early 1980's?” *American Economic Review 90*, 1464-1476.

McLeod, A. I. and Li, W. K. (1983) “Diagnostic Checking ARMA Time Series Models Using Squared-Residual Autocorrelations” *Journal of Time Series Analysis 4*, 269-73.

Mizrach, B. (1991) “A Simple Nonparametric Test for Independence.” Unpublished manuscript.

Morley, J. and J. Piger (2004) “The Importance of Nonlinearity in Reproducing Business Cycle Features” Unpublished manuscript.

Nychka, D., Ellner, S., Gallant, A.R., and McCaffrey, D. (1992) “Finding Chaos in Noisy Systems.” *Journal of the Royal Statistical Society B 54*, 399-426.

Palm, F. C. and Pfann, G. A. (1997) “Sources of Asymmetry in Production Factor Dynamics” *Journal of Econometrics 82*, 361-92.

Patterson, D. M. and Ashley, R. (2000). *A Nonlinear Time Series Workshop.* Kluwer:Norwell .

Potter, S. M. (1995) “A Nonlinear Approach to U.S. GNP” *Journal of Applied Econometrics 10*, 109-125.

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

Ramsey, J. B. and Rothman, P. (1996) "Time Irreversibility and Business Cycle Asymmetry," *Journal of Money, Credit, and Banking 28*, 1-21.

Ramsey, J.B. (1969) "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis." *Journal of the Royal Statistical Society B 31*, 350-371.

Saikkonen, P. and Luukkonen (1988) "Lagrange Multiplier Tests for Testing Nonlinearities in Time Series Models" *Scandinavian Journal of Statistics 15*, 55-68.

Subba Rao, T. and Gabr, M. (1980) "A Test for Linearity of Stationary Time Series Analysis" *Journal of Time Series Analysis 1*, 145-58.

Teräsvirta, T. and H. Anderson (1992) "Characterising Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models," *Journal of Applied Econometrics 7*, 119-36.

Teräsvirta, T. (1994) "Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models," *Journal of the American Statistical Association 89*, 208-218.

Tong, H. (1983) *Threshold Models in Non-linear Time Series Analysis*, Springer-Verlag: New York.

Tsay, R. S. (1986) "Nonlinearity Tests for Time Series" *Biometrika 73,* 461-6.

Tsay, R. S. (1991) "Detecting and Modeling Nonlinearity in Univariate Time Series" *Statistica Sinica 1*, 431-452.

Verbrugge, R. (1997) "Investigating Cyclical Asymmetries" *Studies in Nonlinear Dynamics and Econometrics 2*, 15-22.

White, H. (1989) "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models." *Journal of the American Statistical Association 84*, 1003-1013.

Appendix 1

Nonlinearity Tests Considered

McLeod-Li Test

This test for ARCH effects was proposed by McLeod and Li (1983) based on a suggestion in Granger and Andersen (1978). It looks at the autocorrelation function of the squares of the prewhitened data and tests whether corr(x_t^2 , x_{t-j}^2) is non-zero for some j. The autocorrelation at lag j for the squared residuals $\{x_t^2\}$ is estimated by:

$$\hat{r}(j) = \frac{\sum_{t=1}^{N}\left(x_t^2 - \hat{\sigma}^2\right)\left(x_{t-j}^2 - \hat{\sigma}^2\right)}{\sum_{t=1}^{N}\left(x_t^2 - \hat{\sigma}^2\right)} \qquad \text{where} \qquad \hat{\sigma}^2 = \sum_{t=1}^{N}\frac{x_t^2}{N}.$$

Under the null hypothesis that x_t is an i.i.d process McLeod and Li (1983) show that, for sufficiently large L,

$$Q = N(N+2)\sum_{j=1}^{L}\frac{\hat{r}^2(j)}{N-j}$$

is asymptotically $\chi^2(L)$ under the null hypothesis of a linear generating mechanism for the data. Typically L is taken to be around 20; below results are quoted for L = 24.

Engle LM Test

This test was proposed by Engle (1982) to detect ARCH disturbances; as Bollerslev (1986) suggests, it should also have power against GARCH alternatives. As with most Lagrange Multiplier tests, the test statistic itself is based on the R^2 of an auxiliary regression, in this case:

$$x_t^2 = \alpha_o + \sum_{i=1}^{M} \alpha_i x_{t-i}^2 + \nu_t .$$

Under the null hypothesis of a linear generating mechanism for x_t, NR^2 for this regression is asymptotically distributed $\chi^2(M)$. Below results are quoted for M = 5.

BDS Test

The BDS test is a nonparametric test for serial independence based on the correlation integral of the scalar series, $\{x_t\}$. For embedding dimension m, let $\{\mathbf{X}_t^m\}$denote the sequence of *m*-histories generated by $\{x_t\}$: $\{\mathbf{X}_t^m\} = (x_t, \ldots x_{t+m-1})$.

Then the correlation integral $C_{m,T}(\epsilon)$ for a realization of N is given by:

$$C_{m,N}(\epsilon) = \sum_{t<s} I_\epsilon\left(X_t^m, X_s^m\right) \left\{\frac{2}{N_m(N_m - 1)}\right\}$$

where $N_m = N - (m - 1)$ and $I_\epsilon(X_t^m , X_s^m)$ is an indicator function which equals one if the sup norm $\|X_t^m - X_s^m\| < \epsilon$ and equals 0 otherwise. Basically, $C_{m,N}(\epsilon)$ counts up the number of m-histories that lie within a hypercube of size ϵ of each other. Brock, Dechert, and Scheinkman (1996) exploit the asymptotic normality of $C_{m,N}(\epsilon)$ under the null hypothesis that $\{x_t\}$ is an i.i.d. process to obtain a test statistic which asymptotically converges to a unit normal. This convergence requires extremely large samples for values of the embedding dimension (m) much larger than 2, so attention here is restricted to the cases m = 2, 3, and 4. Where (as here) the data has been normalized to unit variance, the test is ordinarily computed for ϵ =.5, 1, and 2; results are quoted below for $\epsilon = 1$.

According to de Lima (1997) the BDS test requires no moment restrictions. Apparently, this follows from the fact that the test maps the sup norm $\|X_t^m - X_s^m\|$ onto [0, 1] in $\mathbb{R}^1$. However, de Lima (1997) also points out that the existence of the second moment is probably required when the test is applied (as it must be) to the residuals from a linear regression.

Tsay Test

The Tsay (1986) test is a generalization of the Keenan (1985) test; it explicitly looks for quadratic serial dependence in the data, using quadratic terms lagged up to k periods.

Let the K = k(k+1)/2 column vectors $V_1 \ldots V_K$ contain all of the unique crossproducts of the form $x_{t-i}\, x_{t-j}$, where $i \in [1, k]$ and $j \in [i, k]$. Thus, $v_{t,1} = x^2_{t-1}$, $v_{t,2} = x_{t-1}\, x_{t-2}$, $v_{t,3} = x_{t-1}\, x_{t-3}$, ... $v_{t,k} = x_{t-1}\, x_{t-k}$, $v_{t,k+1} = x^2_{t-2}$, $v_{t,k+2} = x_{t-2}\, x_{t-3}$, $v_{t,k+3} = x_{t-2}\, x_{t-4}$, ... and $v_{t,K} = x^2_{t-k}$. And let $\hat{v}_{t,i}$ denote the projection of $v_{t,i}$ on the subspace orthogonal to $x_{t-1}, \ldots, x_{t-k}$ – i.e., the residuals from a regression of $v_{t,i}$ on $x_{t-1}, \ldots, x_{t-k}$.

The parameters $\gamma_1 \ldots \gamma_K$ are then estimated by applying OLS to the regression equation

$$x_t = \gamma_o + \sum_{i=1}^{K} \gamma_i \hat{v}_{t,i} + \eta_t.$$

A value of k = 5 is used below, so that K = 15. The Tsay test statistic is then just the usual F statistic for testing the null hypothesis that $\gamma_1 \ldots \gamma_K$ are all zero.

Hinich Bicovariance Test

This test assumes that $\{x_t\}$ is a realization from a third-order stationary stochastic process and tests for serial independence using the sample bicovariances of the data. The (r,s) sample bicovariance is defined as:

$$C_3(r,s) \quad = \quad (N - s)^{-1} \sum_{t=1}^{N-s} x_t\, x_{t+r}\, x_{t+s} \qquad\qquad 0 \le r \le s.$$

The sample bicovariances are thus a generalization of a skewness parameter. The $C_3(r,s)$ are all zero for zero mean, serially i.i.d. data. One would expect non-zero values for the $C_3(r,s)$ from data in which x_t depends on lagged crossproducts, such as $x_{t-i}x_{t-j}$ and higher order terms.

Let $G(r,s) = (N - s)^{½} C_3(r,s)$ and define X_3 as

$$X_3 \quad = \quad \sum_{s=2}^{\ell} \sum_{r=1}^{s-1} [G(r,s)]^2.$$

Under the null hypothesis that $\{x_t\}$ is a serially i.i.d. process, Hinich and Patterson (1995) show that X_3 is asymptotically distributed $\chi^2(\ell\ [\ell -1]/2)$ for $\ell < N^{½}$; based on their simulations, they recommend using $\ell = N^{.4}$. The X_3 statistic detects non-zero third order correlations; it can be considered a generalization of the Box-Pierce portmanteau statistic.

Hinich Bispectrum Test

This nonparametric test examines the third order moments (bicovariances) of the data in the frequency domain to obtain a direct test for a nonlinear generating mechanism, irrespective of any linear serial dependence which might be present. Consequently, when this test rejects, one need not worry about the possibility that the linear prewhitening model has failed to remove all linear serial dependence in the data. More importantly for the present context, this test's sole focus on nonlinear serial dependence implies that it is making substantially different use of the sample bicovariance data than does the Hinich Bicovariance test described above.

Suppose that $\{y_t\}$, the series of interest, is a third-order stationary time series with, for expositional convenience, $E\{y_t\} = 0$. The series $\{y_t\}$ might be serially correlated, in which case it is distinct from the prewhitened fitting error series denoted $\{x_t\}$ above. Letting $c_{yyy}(r,s)$ denote the third order cumulant function for $\{y_t\}$,

$$c_{yyy}(r,s) \quad = \quad E\left[y_t\, y_{t+r}\, y_{t+s}\right],$$

the bispectrum of $\{y_t\}$ at frequency pair (f_1, f_2) is its (double) Fourier transform:

$$B_y(f_1,f_2) \quad = \quad \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} c_{yyy}(r,s)\ \exp[-i2\pi(f_1 r + f_2 s)].$$

$B_y\,(f_1, f_2)$ is a spatially periodic function of (f_1, f_2), whose principal domain is the triangular set $\Omega = \{0 < f_1 < \frac{1}{2},\ f_2 < f_1,\ 2f_1 + f_2 < 1\}$; see Brillinger and Rosenblatt (1967) for a rigorous treatment of the bispectrum.

If the time series $\{y_t\}$ is linear – so that it can be expressed as

$$y_t \quad = \quad \sum_{n=0}^{\infty} a(n)\ u(t-n)\,,$$

where $u_t \sim \text{iid}(0, \sigma^2)$ and the sequence of weights $\{a(n)\}$ is fixed – then the square of the skewness function

$$\Psi^2(f_1,f_2) \equiv \frac{|B_y(f_1,f_2)|^2}{S_y(f_1)\,S_y(f_2)\,S_y(f_1+f_2)} = \frac{\mu_3^2}{\sigma^6}$$

is a constant for all frequency pairs (f_1, f_2) in Ω, where $S_y(f)$ is the spectrum of $\{y_t\}$ at frequency f. This result was first proven in Brillinger and Rosenblatt (1967); an elementary proof is given in Hinich (1982).

Under the null hypothesis of a linear generating mechanism, this result implies that the sample estimates of Ψ^2 (f_1, f_2) for different frequency pairs will differ from one another no more than one would expect due to sampling error. In particular, Hinich (1982) shows that consistent sample estimates of $2\Psi^2$ (f_1,f_2) are asymptotically distributed as a noncentral chi-squared variate, $\chi^2(2, \lambda)$, with constant non-centrality parameter (λ) under the null hypothesis of linearity; whereas, if the null hypothesis of linearity is false, then λ is dependent on f_1 and f_2. The Hinich test then uses an expression for the asymptotic distribution of the interdecile range of observations from a specified distribution given by David (1970) to test whether the dispersion in the estimates of $2\Psi^2$ (f_1,f_2) exceeds that which one would expect under the null hypothesis. The interdecile (rather than the interquartile) range is used here because it yields test results which are more robust to non-gaussianity in the data. (See also Subba Rao and Gabr (1980) for an earlier approach.)

The bispectrum $\{B_y$ (f_1, f_2)$\}$ is consistently estimated using an average of appropriate triple products of the Fourier representation of the observed time series. This average is taken over a square containing M adjacent frequency pairs. As with smoothing of a periodogram so as to obtain a consistent estimate of the spectrum, large M reduces the variance of the estimator at the cost of introducing some small sample bias. Hinich (1982) shows that M must exceed $N^{.5}$ in order to consistently estimate B_y (f_1, f_2); based on simulation results in Ashley, Patterson, and Hinich (1986) M, is set to the integer closest to $N^{.6}$ in the calculations reported below.

The Hinich Bispectrum test has the nice property that it is unaffected by the application of a linear filter to y_t. (This follows from the fact that the squared skewness function, Ψ^2 (f_1,f_2), is invariant to linear filtering; see Ashley, Patterson, and Hinich (1986, p. 174) for a proof.) Consequently – and in contrast to other approaches – the results from the Hinich Bispectrum test are robust to any errors one might make in pre-whitening the sample data.

Appendix 2

The Sizes of the Nonlinearity Tests

Like most econometric procedures, the tests described in Appendix 1 are only asymptotically justified. Particular concern has been expressed about the validity of the BDS test for reasonable sample sizes and addressed, to some degree, in Brock, et al. (1991). More recently, de Lima (1997) has considered the behavior of a number of nonlinearity tests where the moment restriction assumptions underlying the asymptotic distributions of these tests are not satisfied, finding particular problems in situations involving leptokurtic (heavy-tailed) data.

Because we share these concerns, we routinely bootstrap the significance levels of all the tests used here. This is very straightforward. After pre-whitening, so that the data is (asymptotically) serially i.i.d. under the null hypothesis of a linear generating mechanism, we draw 1000 N-samples at random from the empirical distribution of the observed N-sample of data – i.e., from the fitting errors of the estimated AR(p) model. The bootstrap significance level for a given test is then just the fraction of these 1000 "new" N-samples for which the test statistic exceeds that observed in the sample data. It is simple enough to confirm that 1000 bootstrap replications is sufficient by merely observing that the results are invariant to increasing this number; it is distinctly less clear that N itself is sufficiently large: after all, the pre-whitening procedure and bootstrap itself are themselves only asymptotically justified.

Consequently, it is of interest to examine the actual size of each test for samples of length similar to that used in the models for U.S. real output examined in Section 4. To that end, 200 serially i.i.d. variates were generated from each of four distributions: gaussian, exponential, Student's t with 5 degrees of freedom, and the symmetric stable Paretian distribution with index

$\alpha = 1.93$. The exponential distribution is quite asymmetric. Both of the latter two distributions are heavy-tailed – to the point where the variance does not exist for the symmetric stable Paretian distribution with this index value. (Symmetric stable Paretian variates have finite variance only for $\alpha \geq 2.00$; the value $\alpha = 1.93$ used here is Fama's (1965) estimate for U.S. stock price data. The Paretian variates were obtained using the exact algorithm given by Kanter and Steiger (1974).) Since the bootstrap is actually applied to the AR(p) fitting errors, we also examined the actual sizes of the tests for linearly dependent data, where the observations are generated by an AR(2) process driven by innovations generated from each of these distributions. The AR(2) process used was $y_t = .28y_{t-1} + .08y_{t-2} + \epsilon_t$ or, equivalently, $y_t = (1 - .456B)(1 + .176B)y_t + \epsilon_t$.

The results of these calculations are given in Table 7 below. Under the null hypothesis that the actual size is .05, an (asymptotic) 95% confidence interval for these estimates is (.036, .064); results significantly different from .05 are shown in bold. All figures quoted are based on 1000 samples, each of length 200, bootstrapped from the fitting errors of an AR(p) pre-whitening model, where p is chosen (for each sample) to minimize the value of the BIC. The parameters L, p, m, k, ℓ , and M are defined in Appendix 1, where each test is discussed. BDS test results were calculated for the parameter ϵ equal to one half, one, and two standard deviations; for brevity, results are quoted only for $\epsilon = 1$.

We observe that the actual sizes for these bootstrapped tests appear to be satisfactory in all cases except the BDS test with embedding dimension (m) exceeding two. (The fact that several of the size estimates not involving the BDS test lie outside the 95% confidence interval around .05 is inconsequential in view of the number of estimates made.) Consequently, BDS test results are only quoted only for embedding dimension two in the remainder of the paper. These BDS size results differ from those given in Patterson and Ashley (2000, Tables 4-1, 4-2) in that it was artificially assumed there that the correct value of p for the AR(p) pre-whitening model was known. Since the BDS test is correctly sized at all three embedding dimensions in those results, the size problem here at larger embedding dimensions is evidently due to the BDS test's high sensitivity to the minor amounts of linear dependence remaining in the data on those occasions where the prewhitening procedure mis-identifies the order of the AR(p) prewhitening model.

We conclude that it is reasonable to proceed using the bootstrapped tests for samples of roughly this length or larger without further concern about moment restrictions or the form of the data's distribution.

Table 7
Empirical Size of 5% Tests

	McLeod-Li	Engle LM	BDS			Tsay	Bicov.	Bispectrum
	L = 24	p = 5	m = 2	m = 3	m = 4	k = 5	ℓ = 8	M = 24
Serially i.i.d. Data								
Gaussian	.045	.056	.057	**.075**	**.088**	.053	.052	.048
Student's t(5)	.045	.044	.056	.057	**.076**	.062	.061	.051
Exponential	.037	**.070**	.057	.061	**.067**	.058	.061	.053
Paretian α =1.93	.048	**.032**	.062	.062	**.065**	.062	.049	.053
Linearly Dependent {AR(2)} Data								
Gaussian	.043	.060	.044	.046	.052	.057	.047	.046
Student's t(5)	.045	.056	**.071**	**.100**	**.121**	.055	.056	.047
Exponential	**.026**	.045	.062	.064	**.071**	.047	.059	.047
Paretian α =1.93	.037	.041	.055	**.068**	**.072**	.045	.049	.059