JOURNAL OF APPLIED ECONOMETRICS, VOL. 7, S61–S82 (1992)

THE LIKELIHOOD RATIO TEST UNDER NONSTANDARD CONDITIONS: TESTING THE MARKOV SWITCHING MODEL OF GNP

B. E. HANSEN
University of Rochester, Harkness Hall, Rochester, New York 14627, USA

SUMMARY

A theory of testing under non-standard conditions is developed. By viewing the likelihood as a function of the unknown parameters, empirical process theory enables us to bound the asymptotic distribution of standardized likelihood ratio statistics, even when conventional regularity conditions (such as unidentified nuisance parameters and identically zero scores) are violated. This testing methodology is applied to the Markov switching model of GNP proposed by Hamilton (1989). The standardized likelihood ratio test is unable to reject the hypothesis of an AR(4) in favour of the Markov switching model. Instead, we find strong evidence for an alternative model. This model, like Hamilton's, is characterized by parameters which switch between states, but the states arrive independently over time, rather than following an unrestricted Markov process. The primary difference, however, is that the second autoregressive parameter, in addition to the intercept, switches between states.

1. INTRODUCTION

Applied econometrics is increasingly dominated by nonlinear models and estimation techniques. The absence of a body of finite sample theory for nonlinear models means that applied research must rely either on asymptotic theory or bootstrapping for inference. The primary asymptotic distributional theory for nonlinear models runs roughly as follows. In a sufficiently large sample the estimator nears the true parameter vector. Via a Taylor's expansion the parameter estimates are equal to their true value, plus the score evaluated at the true value, divided by the second derivative matrix evaluated at median points.[1] The likelihood surface is assumed to be approximately quadratic in this region, so the second derivatives are approximately constant (that is, not a function of the parameters). Since the score is mean zero, if it has positive variance, we can apply a central limit theorem, and conclude that the estimator has an asymptotic multivariate normal distribution.

There appear to be two key assumptions to this argument. First, the likelihood surface must be locally quadratic. We must interpret 'locally' to mean that the likelihood surface is approximately quadratic over the region in which both the null hypothesis and the global optimum lie (with high probability).[2] In fact this condition is routinely violated in many applications. For example, if some parameters are not identified under the null hypothesis, then

[1] The rows of the matrix are not necessarily evaluated at the same points.

[2] The requirement that the global optimum lie in the locally quadratic region 'with high probability' is somewhat circular, since the argument is designed to provide a *distributional* theory. The conventional proof circumvents this problem by appealing to the consistency of the estimator.

0883–7252/92/0S0S61–22$16.00

Received May 1991
Revised August 1992

the likelihood function is *flat* (with respect to the unidentified parameters) at the optimum. In other cases the likelihood surface has more than one local optima, and the null hypothesis may not lie on the same 'hill' as the global optimum. In this case the likelihood surface is far from quadratic in the region between the global optimum and the null hypothesis. The second key assumption is that the score must have a positive variance. This condition is violated when the score is identically zero under the null hypothesis, which occurs when the null hypothesis yields a local maximum, minimum, or inflection point. Some of these problems have been outlined in the literature before, and separate methods proposed for 'handling' the distributional theory in these special cases.

Davies (1977, 1987) analyses the problem of unidentified nuisance parameters. He suggests viewing the test statistic as a function of the nuisance parameter, in order to apply empirical process theory. Davies bounds the maximum of the empirical process using a crossing-point argument. Hansen (1991) extends the empirical process theory to a wider class of estimation problems and test statistics, but instead of bounding the maximum, provides a direct method to compute critical values, using the empirical covariance function of the empirical process.

Lee and Chesher (1986) study the Lagrange multiplier (LM) test in the case of identically zero scores. They suggest examining higher-order derivatives at the null. This may be useful if the higher-order derivatives are also not identically zero; but even if they are not, the power of their test is not clear. For example, this test will have asymptotic zero power if the likelihood attains a local maximum at the null.

Each of the above papers present methods which are useful in certain special cases. No general results appear to exist. In an attempt to fill this void, this paper takes a new approach to testing which does not require either that the likelihood be locally quadratic or that the scores (or any other derivative) have positive variance. We work directly with the likelihood surface, viewing the likelihood function as an empirical process of the unknown parameters. Empirical process theory is used to derive a bound for the asymptotic distribution of a standardized likelihood ratio statistic. The distribution depends upon the covariance function of the empirical process associated with the likelihood surface, but we show that the distribution of this empirical process can be easily obtained via simulation.

To my knowledge no-one has proposed a similar methodology before. An analogous bound, however, was proposed by Horowitz and McAleer (1989) in the context of non-nested hypothesis testing. Bounds for non-standard Wald tests have also been analysed in Kemp (1991).

This new testing apparatus is set to work on the Markov switching model of output proposed by Hamilton (1989). Hamilton modelled postwar US GNP growth rates as the sum of an AR(4) process and a Markov process. This may be interpreted as a model where one of the parameters (the mean) switches between two values according to a Markov transition process. Hamilton argued that this model was a better description of the data than the traditional AR model with a fixed mean. As recognized in his original paper, however, this model is plagued by not just one, but *all* of the problems mentioned above. Two nuisance parameters (the transition probabilities) are not identified under the null hypothesis. The null hypothesis also yields a local optimum of the likelihood surface, and higher-order derivatives also appear to be zero. This yields a singular information matrix under the null. Being highly nonlinear, the model produces numerous local optima as well. Recognizing the inapplicability of standard theory, Hamilton (1989) did not attempt a formal hypothesis test of the null of an AR(4) versus his Markov switching model.

The standardized LR test, which is a valid statistical test to discriminate between these models, fails to reject the null of an AR(4) in favour of the Markov switching model.

Apparently, the presence of the two nuisance parameters gives the likelihood surface sufficient freedom so that we cannot reject the possibility that the apparent 'significant' coefficients could simply be due to sampling variation.

A series of Monte-Carlo studies are also presented. It is found that the test has virtually no size distortion in this application. Thus the failure to reject the null appears not to be a consequence of the use of a bound for the asymptotic distribution. The simulations also reveal that the power of the test is quite good, especially when no autoregressive component is included in the estimated specification.

Hamilton's Markov switching model, however, is quite restrictive in only allowing one parameter to vary with the Markov state. We find strong evidence for an alternative model, in which growth rates are modelled as an AR(2), with the intercept and second AR parameter varying between states. Further, there is no persistence in the states, as the model accepts the restriction that the probability of being in one state or the other is independent of the current state. That is, we find that GNP is characterized by a *simple switching* model, rather than a *Markov switching* model for GNP. The standardized LR test rejects the null hypothesis of an AR(4) in favour of this alternative switching model around the 1 per cent level.

Section 2 presents the main theoretical results in a simplified environment without nuisance parameters. Technical details are de-emphasized in favour of intuition. Section 3 outlines the theory more completely, allowing for nuisance parameters (both identified and non-identified). These sections develop the apparatus to analyse the likelihood function as an empirical process. Sections 4 and 5 use these methods to analyse postwar quarterly US GNP. Section 4 analyses Hamilton's Markov switching model, and section 5 proposes an alternative simple switching model. A conclusion follows.

Concerning notation, the symbol '$\Rightarrow$' is used to denote weak convergence of probability measures with respect to the uniform metric, and '$\|\cdot\|$' is used to denote the Euclidean metric. All limits are taken as the sample size, n, tends to positive infinity.

2. THE LIKELIHOOD SURFACE AS AN EMPIRICAL PROCESS

Let us start with a relatively simple problem. Take a likelihood function which is a function of an unknown parameter $\alpha \in A$ where A is some compact metric space. Suppose that the log-likelihood can be written in the form:

$$L_n(\alpha) = \sum_{i=1}^{n} l_i(\alpha),$$

with the null and alternative hypotheses:

$$H_0\colon \quad \alpha = \alpha_0, \qquad H_1\colon \alpha \neq \alpha_0.$$

It will be very useful to define the likelihood ratio (LR) function:

$$\mathrm{LR}_n(\alpha) = L_n(\alpha) - L_n(\alpha_0) = \sum_{i=1}^{n} [l_i(\alpha) - l_i(\alpha_0)].$$

This function gives the sequence of Neyman–Pearson likelihood ratio test statistics for the test of the null against each simple alternative hypothesis. This is not a function which is commonly used, but it has a number of useful features. Since the likelihood ratio surface is simply a level-shift of the likelihood surface, the maximum likelihood estimator (MLE) is given by the parameter value which maximizes the likelihood ratio surface. It follows as well that the likelihood ratio test statistic for H_0 against H_1 is given by the supremum of the likelihood ratio

surface:

$$\mathrm{LR}_n = \sup_{\alpha \in A} \mathrm{LR}_n(\alpha).$$

The LR surface can be decomposed into its mean, and deviation from mean:

$$\mathrm{LR}_n(\alpha) = R_n(\alpha) + Q_n(\alpha) \tag{1}$$

where

$$R_n(\alpha) = E[\mathrm{LR}_n(\alpha)],$$

is the mean, and

$$Q_n(\alpha) = \sum_1^n q_i(\alpha),$$

is the deviation from the mean, where

$$q_i(\alpha) = [l_i(\alpha) - l_i(\alpha_0)] - E[l_i(\alpha) - l_i(\alpha_0)].$$

It is useful to reflect upon decomposition (1). Under standard regularity conditions, $n^{-1}R_n(\alpha) \to_p R(\alpha)$ for all α, where $R(\alpha) = E[l_i(\alpha) - l_i(\alpha_0)]$. The function $R_n(\alpha)$ is maximized precisely at the true parameter vector (which is α_0 under the null). It follows that under the null hypothesis, $R_n(\alpha)$ is nonpositive, and strictly negative for $\alpha \neq \alpha_0$.

If the econometrician could actually observe $R_n(\alpha)$ there would be no uncertainty. In the real world, however, an econometrician observes $\mathrm{LR}_n(\alpha)$, which contains the influence of the random function $Q_n(\alpha)$. Indeed, the existence of random fluctuations in the function $Q_n(\alpha)$ is why the likelihood is maximized at some value of α other than α_0. We can therefore find some insight into the behaviour of the optimization problem by studying the stochastic process $Q_n(\alpha)$.

When properly standardized, we find

$$\frac{1}{\sqrt{n}} Q_n(\alpha) = \frac{1}{\sqrt{n}} \sum_1^n q_i(\alpha) \Rightarrow Q(\alpha), \tag{2}$$

where $Q(\alpha)$ is a mean zero Gaussian process with covariance function

$$K(\alpha_1, \alpha_2) = E[q_i(\alpha_1) q_i(\alpha_2)]. \tag{3}$$

The empirical process result (2) is a natural generalization of the classical central limit theorem. For each value of α, $Q(\alpha)$ is a normal random variable with mean zero and variance $K(\alpha, \alpha)$. The function $K(\cdot, \cdot)$ describes the covariances between $Q(\alpha)$ at different values of α.

The decomposition (1) can be rewritten as an asymptotic approximation:

$$\begin{aligned} \frac{1}{\sqrt{n}} \mathrm{LR}_n(\alpha) &= \frac{1}{\sqrt{n}} R_n(\alpha) + \frac{1}{\sqrt{n}} Q_n(\alpha) \\ &= \frac{1}{\sqrt{n}} R_n(\alpha) + Q(\alpha) + o_p(1) \end{aligned} \tag{4}$$

where the $o_p(1)$ term holds uniformly in α. (4) states that the LR surface equals (in large samples) the mean function plus a Gaussian process. The Gaussian process $Q(\alpha)$ is completely determined by the covariance function $K(\cdot)$ in (3), which can be estimated from the data (we will discuss this in section 3.2). The mean function, $R_n(\alpha)$ is unknown. Standard asymptotic theory requires that $R_n(\alpha)$ is well-behaved. We can avoid this requirement by instead appealing

to the fact that $R_n(\alpha) \leqslant 0$ for all α when the null hypothesis is true. This gives

$$\frac{1}{\sqrt{n}} \mathrm{LR}_n(\alpha) \leqslant \frac{1}{\sqrt{n}} Q_n(\alpha) \Rightarrow Q(\alpha). \tag{5}$$

From (5), we can find a bound for the asymptotic distribution of the standard LR test of H_0 against H_1. Since $\mathrm{LR}_n = \sup_\alpha \mathrm{LR}_n(\alpha)$, we have as $n \to \infty$,

$$P\left\{\frac{1}{\sqrt{n}} \mathrm{LR}_n \geqslant x\right\} \leqslant P\left\{\sup_\alpha \frac{1}{\sqrt{n}} Q_n(\alpha) \geqslant x\right\} \to P\left\{\sup_\alpha Q(\alpha) \geqslant x\right\}. \tag{6}$$

While an interesting theoretical observation, it is not clear that (6) provides a distributional bound which is very useful in practice. The process $Q(\alpha)$ is Gaussian, but it it not standardized. As $\alpha \to \alpha_0$, for example, $R_n(\alpha)$ and $Q_n(\alpha)$ vanish. Since the MLE of α will be converging in probability to α_0, $\mathrm{LR}_n(\alpha)$ will be maximized at a value of α which is 'close' to α_0 in large samples. There is no such requirement upon $Q(\alpha)$, however. Thus the bound $\sup_\alpha Q(\alpha)$ will be over-conservative in practice. In fact, it can be shown that this test will have true size which converges to zero as the sample size diverges, even though the test will reject with probability one (asymptotically) under the alternative.

A sensible alternative is to standardize the likelihood ratio so that all values of α yield the same variance. This will preserve the main features of the likelihood surface under the alternative hypothesis, but reduce the over-conservative tendency under the null.

We start with the variance function associated with the covariance function:

$$V(\alpha) = K(\alpha, \alpha).$$

Consider the sample analogue,

$$\begin{aligned} V_n(\alpha) &= \sum_{i=1}^{n} \left[l_i(\alpha) - l_i(\alpha_0) - \frac{1}{n} \mathrm{LR}_n(\alpha) \right]^2 \\ &= \sum_{i=1}^{n} \left[l_i(\alpha) - l_i(\alpha_0) \right]^2 - \frac{1}{n} \mathrm{LR}_n(\alpha)^2. \end{aligned}$$

We then have (leaving aside technical details until section 3):

$$Q_n^*(\alpha) = \frac{Q_n(\alpha)}{V_n(\alpha)^{1/2}} \Rightarrow \frac{Q(\alpha)}{V(\alpha)^{1/2}} = Q^*(\alpha), \text{ say.}$$

The Gaussian process $Q^*(\alpha)$ has unit variance for all $\alpha \neq \alpha_0$, and thus $Q^*(\alpha) \equiv N(0, 1)$. Now for $\alpha \neq \alpha_0$, define the standardized likelihood ratio process:

$$\mathrm{LR}_n^*(\alpha) = \frac{\mathrm{LR}_n(\alpha)}{V_n(\alpha)^{1/2}},$$

and the standardized likelihood ratio statistic:

$$\mathrm{LR}_n^* = \sup_\alpha \mathrm{LR}_n^*(\alpha)$$

We now conclude our discussion with the bound:

$$\begin{aligned} P\{\mathrm{LR}_n^* \geqslant x\} &\leqslant P\{\sup_\alpha Q_n^*(\alpha) \geqslant x\} \\ &\to P\{\sup_\alpha Q^*(\alpha) \leqslant x\} = F^*(x). \end{aligned}$$

As we show later, we can obtain good approximations to the distribution function $F^*(x)$. One negative feature of the above asymptotic distribution is that it is a *bound*. Thus tests based on this approach may be conservative (under-rejection when the null is true), and hence suffer a loss in effective power (ability to reject the null when it is false). Simulations reported in sections 4 and 5, however, suggest that the test is not conservative in the applications considered in this paper.

In some cases the inequality will be an equality, eliminating this concern. Consider the simple location model: y_t i.i.d. $N(\alpha, 1)$ with $H_0: \alpha = 0$ vs. $H_1: \alpha > 0$. Here

$$\mathrm{LR}_n(\alpha) = \frac{1}{2}\sum_1^n [(y_t - \alpha)^2 - y_t^2] = n[\alpha\bar{y} - \alpha^2/2].$$

So

$$V_n(\alpha) = \frac{1}{n}\sum_1^n [\alpha y_t - \alpha^2/2 - (\alpha\bar{y} - \alpha^2/2)]^2 = \alpha^2\hat{\sigma}_y^2$$

and thus

$$\mathrm{LR}_n^*(\alpha) = \sqrt{n}(\bar{y} - \alpha/2)/\hat{\sigma}_y.$$

We find that

$$\mathrm{LR}_n^* = \sup_\alpha \mathrm{LR}_n^*(\alpha) = \sqrt{n}\bar{y}/\hat{\sigma}_y$$

which is the standard t-statistic for the test of H_0 against H_1. In this simple example the standardized LR statistic has a conventional interpretation and distribution. This will not always be the case, but it suggests that the structure of the standardized LR statistic is not as unconventional as appears at first glance.

3. GENERAL THEORY

3.1. Allowing for Nuisance Parameters

The previous section was meant to be motivational, since most problems of interest contain nuisance parameters. Suppose that the model has log-likelihood

$$L_n(\beta, \gamma, \theta) = \sum_{i=1}^n l_i(\beta, \gamma, \theta)$$

with parameter vectors $\beta \in \mathrm{B}$, $\gamma \in \Gamma$, and $\theta \in \Theta$. The hypothesis takes the form

$$H_0: \beta = 0 \qquad H_1: \beta \neq 0.$$

Note that θ and γ are nuisance parameters. Assume that θ is fully identified, but γ is not identified under H_0. (This requires that $L_n(0, \gamma, \theta)$ not depend upon γ.) In order to apply a testing method similar to that suggested in section 2, we have to eliminate the parameter vector θ. We do this by concentration.

Set $\alpha = (\beta', \gamma')$, $\mathrm{A} = \mathrm{B} \times \Gamma$, and $L_n(\alpha, \theta)$ and $l_i(\alpha, \theta)$ accordingly. Define the sequence of parameter estimates

$$\hat{\theta}(\alpha) = \max_{\theta \in \Theta} L_n(\alpha, \theta) \tag{7}$$

which are the maximum-likelihood estimates of θ for fixed values of α. The concentrated

likelihood function is then

$$\hat{L}_n(\alpha) = L_n(\alpha, \hat{\theta}(\alpha)).$$

Ideally, we would like to be working with the large-sample concentrated likelihood function given by

$$L_n(\alpha) = L_n(\alpha, \theta(\alpha))$$

where

$$\theta(\alpha) = \underset{\theta \in \Theta}{\text{Argmax}} \lim_{n \to \infty} \frac{1}{n} EL_n(\alpha, \theta)$$

is the pseudo-true value of θ, for fixed α. In order for the concentration argument to work, we require that $\hat{\theta}(\alpha)$ is consistent for $\theta(\alpha)$ at rate $\sqrt{n}$, uniformly in α. Set $D(\alpha) = \hat{\theta}(\alpha) - \theta(\alpha)$. Formally, we assume

$$\sup_{\alpha \in A} \sqrt{n} \,\| D(\alpha) \| = O_p(1). \tag{A1}$$

In order to show (A1) from more primitive assumptions, we would have to assume that the maximization problem given in (7) satisfies the standard assumptions for nonlinear estimators. That is, we are assuming that all of the 'trouble' arises in the parameters $\alpha = (\beta', \gamma')$. We further require that the matrix of second derivatives with respect to θ be well behaved. If we define

$$M_n(\alpha, \theta) = \frac{\partial^2}{\partial \theta \, \partial \theta'} L_n(\alpha, \theta),$$

we require

$$\sup_{\alpha \in A, \theta \in \Theta} \| M_n(\alpha, \theta) \| = O_p(n). \tag{A2}$$

By a Taylor's expansion we have

$$L_n(\alpha, \theta(\alpha)) - L_n(\alpha, \hat{\theta}(\alpha)) = D(\alpha)' \frac{\partial}{\partial \theta} L_n(\alpha, \hat{\theta}(\alpha)) + \tfrac{1}{2} D(\alpha)' M_n(\alpha, \theta^*(\alpha)) D(\alpha),$$

where $\theta^*(\alpha)$ lies on a line segment joining $\hat{\theta}(\alpha)$ and $\theta(\alpha)$. This gives

$$\sup_{\alpha \in A} \| L_n(\alpha) - \hat{L}_n(\alpha) \| = \sup_{\alpha \in A} \| D(\alpha)' M_n(\alpha, \theta^*(\alpha)) D(\alpha) \| = O_p(1). \tag{8}$$

We now proceed as in section 2. The likelihood ratio process, its large-sample counterpart, expectation and centred versions are

$$\hat{\mathrm{LR}}_n(\alpha) = \hat{L}_n(\alpha) - \hat{L}_n(0, \gamma),$$

$$\mathrm{LR}_n(\alpha) = L_n(\alpha) - L_n(0, \gamma)$$

$$R_n(\alpha) = E[\mathrm{LR}_n(\alpha)]$$

$$\hat{Q}_n(\alpha) = \hat{\mathrm{LR}}_n(\alpha) - R_n(\alpha)$$

$$Q_n(\alpha) = \mathrm{LR}_n(\alpha) - R_n(\alpha).$$

We could now assume that an empirical process central limit theorem (CLT) holds:

$$\frac{1}{\sqrt{n}} Q_n(\alpha) \Rightarrow Q(\alpha) \tag{9}$$

where $Q(\alpha)$ is a Gaussian process with covariance function

$$K(\alpha_1, \alpha_2) = \lim_{n \to \infty} \frac{1}{n} E[Q_n(\alpha_1) Q_n(\alpha_2)].$$

Set $V(\alpha) = K(\alpha, \alpha)$ to be the associated variance function. Andrews (1991) recently has provided an empirical process CLT under conditions which permit temporal dependence and heterogeneity. Essentially, the likelihood components $l_i(\alpha, \theta(\alpha))$ need to have bounded $2 + \delta$ moments, satisfy a mixing or near epoch dependence condition, and satisfy a smoothness condition with respect to α. It is also usually necessary that the parameter space A be compact, which we shall assume as well. An alternative proof using a bracketing approach has also been provided by Andrews and Pollard (1990).

Using the fact that $R_n(\alpha) = R_n(\beta, \gamma) \leqslant 0$ under the null hypothesis, (8), and (9), we could obtain a limit theory for the concentrated likelihood process:

$$\frac{1}{\sqrt{n}} \hat{\mathrm{LR}}_n(\alpha) \leqslant \frac{1}{\sqrt{n}} \hat{Q}_n(\alpha) = \frac{1}{\sqrt{n}} Q_n(\alpha) + o_p(1) \Rightarrow Q(\alpha).$$

Note that the $o_p(1)$ term holds uniformly in α.

As discussed in the previous section, it appears to make more sense to work with the standardized LR process. Construct the sample variance

$$V_n(\alpha, \hat{\theta}(\alpha)) = \sum_1^n q_i(\alpha, \hat{\theta}(\alpha))^2,$$

where

$$q_i(\alpha, \hat{\theta}(\alpha)) = l_i(\alpha, \hat{\theta}(\alpha)) - l_i(0, \gamma, \hat{\theta}(0, \gamma)) - \frac{1}{n} \hat{\mathrm{LR}}_n(\alpha).$$

The standardized LR function is defined as

$$\hat{\mathrm{LR}}_n^*(\alpha) = \frac{\hat{\mathrm{LR}}_n(\alpha)}{V_n(\alpha)^{1/2}},$$

yielding the standardized LR statistic

$$\hat{\mathrm{LR}}_n^* = \sup_{\alpha \in \mathrm{A}} \hat{\mathrm{LR}}_n^*(\alpha)$$

Define the centred stochastic processes

$$\hat{Q}_n^*(\alpha) = \frac{\hat{Q}_n(\alpha)}{V_n(\alpha)^{1/2}}, \qquad Q_n^*(\alpha) = \frac{Q_n(\alpha)}{V_n(\alpha)^{1/2}}.$$

Instead of assuming that (9) holds, we instead (or additionally) assume that $Q_n^*(\alpha)$ satisfies an empirical process law:

$$Q_n^*(\alpha) \Rightarrow Q^*(\alpha), \tag{A3}$$

where $Q^*(\alpha) = Q(\alpha)/V(\alpha)^{1/2}$ is a Gaussian process with covariance function

$$K^*(\alpha_1; \alpha_2) = \frac{K(\alpha_1, \alpha_2)}{V(\alpha_1)^{1/2} V(\alpha_2)^{1/2}}.$$

(A3) would follow from (9) if $n^{-1} V_n(\alpha)$ converges in probability uniformly to $V(\alpha)$ and we restrict attention to a compact subset of A over which $V(\alpha)$ is uniformly positive definite. This

would exclude consideration of some (arbitrary) neighbourhood of the null hypothesis. We do not need to make this restrictive construction, however, if we alternatively impose the empirical process regularity conditions directly upon $Q_n^*(\alpha)$. In the applications considered in this paper, the likelihood surface is quite smooth with respect to the parameter α, so this is not a problem. This is in contrast to some other models, such as those involving change points or thresholds in which case the likelihood is not smooth with respect to the unidentified parameters.

We find

$$\begin{aligned}\hat{\mathrm{LR}}_n^* &\leqslant \sup_{\alpha \in \mathrm{A}} \hat{Q}_n^*(\alpha) \\ &= \sup_{\alpha \in \mathrm{A}} Q^*{}_n(\alpha) + o_p(1) \\ &\Rightarrow \sup_{\alpha \in \mathrm{A}} Q^*(\alpha) \equiv \mathrm{Sup}\, Q^*.\end{aligned}$$

We have shown the following result.

Theorem, 1. Under (A1)–(A3),

$$P\{\hat{\mathrm{LR}}_n^* \geqslant x\} \leqslant P\{\sup_{\alpha \in \mathrm{A}} \hat{Q}_n^*(\alpha) \geqslant x\} \to \mathrm{P}\{\mathrm{Sup}\, Q^* \geqslant x\}.$$

Theorem 1 provides a bound for the standardized *LR* statistic in terms of the distribution of the random variable Sup Q^*. The assumptions (A1) through (A3) are high-level, but quite weak, in contrast to conventional distributional theory. Thus Theorem 1 is applicable in a much wider class of models than the standard theory. The cost is the presence of the inequality. The fact that the distribution of the test statistic is only bounded means that the test may be conservative and effective power may be lowered. Hence, Theorem 1 should only be used (*vis-à-vis* conventional theory) when it is apparent that the conventional assumptions are invalid.

3.2. Calculating the Asymptotic Distribution

The distribution of the random variable Sup Q^* presented in Theorem 1 is generally non-standard, precluding generic tabulation. Following Hansen (1991), it is quite easy, however, to use the empirical covariance function to generate the asymptotic distribution via simulation.

The random variable Sup Q^* is the supremum of the empirical process $Q^*(\alpha)$, which is completely characterized by its covariance function $K^*(\cdot)$. We do not know K^*, but we have the sample analogue:

$$\hat{K}_n^*(\alpha_1, \alpha_2) = \frac{\sum_1^n q_i(\alpha_1, \hat{\theta}(\alpha_1)) q_i(\alpha_2, \hat{\theta}(\alpha_2))}{V_n(\alpha_1)^{1/2} V_n(\alpha_2)^{1/2}}.$$

Suppose that we can draw i.i.d. Gaussian processes whose covariance function is $\hat{K}_n^*(\cdot, \cdot)$. The supremum of each of these processes (approximately) has the distribution Sup Q^*, where the approximation is only due to the sample discrepancy between $\hat{K}_n^*$ and K^* which vanishes in large samples. Through repeated draws from this urn, we can (approximately) obtain the distribution Sup Q^* from the empirical distribution of the random draws. For example, critical values and p-values can be calculated, and histograms plotted, for any given example.

An easy method to obtain draws from the required family of Gaussian processes is to generate a random sample of $N(0, 1)$ variables $\{u_i\}_1^n$, and then construct

$$\tilde{\mathrm{LR}}^*(\alpha) = \frac{\sum_1^n q_i(\alpha, \hat{\theta}(\alpha)) u_i}{V_n(\alpha)^{1/2}}.$$

It is straightforward to verify that (conditional on the data) the process $\tilde{\mathrm{LR}}^*(\cdot)$ is Gaussian with covariance function $\hat{K}_n^*$. It is also (conditionally) independent from other processes $\tilde{\mathrm{LR}}^*(\cdot)$ constructed with independent samples $\{u_i\}$. It is evident that this construction meets our requirements.

3.3. Practical Issues

The main cost of the procedures advocated here is in the evaluation of the likelihood across different values of $\alpha = (\beta, \gamma)$. The need to concentrate out the identified nuisance parameters (θ) means that, for each value of α, the constrained likelihood needs to be optimized. This can be a major computational burden, even if the parameter space is small.

As far as I can see, the only practical way to evaluate the maximal statistics discussed here is to form a grid search over a relatively small number of values of α. A trade-off arises as a more extensive grid search requires more computation, but reduces the arbitrariness associated with the choice of grid, and may increase the power of the test. For every value of α at which the constrained likelihood is optimized, one needs to calculate only the sequence $\{q_i(\alpha, \hat{\theta}(\alpha))\}$ (an $n \times 1$ vector). From these numbers, both the modified LR statistic and its asymptotic distribution can be calculated.

4. TESTING THE MARKOV SWITCHING MODEL OF GNP

4.1. Testing Hamilton's Markov Switching Model

What is a good univariate model of GNP? Since the degree of persistence in linearly detrended GNP is quite high (seemingly nonstationary), but the amount of persistence in growth rates is relatively low, we will be interested in finding a model for the first difference of the natural log of real GNP, which we will denote by x_t.

A reasonable starting place is the autoregressive (AR) model:

$$\varphi(L)x_t = \mu + e_t, \tag{10}$$

where e_t is i.i.d. perhaps from a normal distribution. The argument for the AR model is that (practically) all covariance stationary processes have an autoregressive representation, which can be written as (10) where the error e_t is white noise. The reasonableness of adopting the AR model is that most of the estimation and inference techniques designed for the AR model are valid under the broader conditions of an AR representation, so an applied researcher need not be worried that he/she has the 'wrong' model. The results of fitting an AR(4) to postwar quarterly US GNP are presented in Table I.[3]

[3] All regressions are reported with heteroscedasticity-consistent standard errors (see White, 1980). Also reported is the Gaussian log-likelihood and the value of the LM test for parameter instability proposed in Nyblom (1989) and Hansen (1990).

Table I. Maximum-likelihood estimates of Gaussian AR model, US real GNP, 1952:2 to 1984:4

Parameter	Estimate	Standard error
μ	0·557	0·140
φ_1	0·310	0·085
φ_2	0·127	0·095
φ_3	−0·121	0·087
φ_4	−0·089	0·090
σ	0·983	0·064
Log-likelihood	−183·669	
LM stability test	0·958	(insignificant at 20 per cent level)

The representation argument does not imply that an AR model is adequate for all purposes. A Gaussian AR model is incompatible, for example, with the observed asymmetry between expansions and contractions. This asymmetry could be 'explained' by an AR model with skewed innovations e_t, but this solution is not completely satisfactory. If, for instant, the errors in the AR representation are not independent, but have conditionally forecastable third moments, then the AR model is suboptimal, since it is not taking into account forecastable asymmetries in the business cycle.

Many alternatives to the AR model are possible. Hamilton (1989) proposed a 'Markov switching' formulation in which a large degree of explanatory power is assigned to the existence of a few 'states' between which the economy shifts according to a Markov process. For GNP growth rates, Hamilton suggested the model

$$x_t = \mu + \mu_d s_t + u_t, \qquad \varphi(L)u_t = e_t, \tag{11}$$

where s_t is a latent dummy variable equalling either 1 or 0. The transitions between these states are governed by the transition probabilities

$$P\{s_t = 1 \mid s_{t-1} = 1\} = p$$

$$P\{s_t = 0 \mid s_{t-1} = 0\} = q.$$

In his paper, Hamilton set the autoregressive order equal to four. In order to estimate the model by maximum-likelihood, Hamilton added the assumption that e_t is i.i.d. $N(0, \sigma^2)$ and independent of $\{s_t\}$.

Table II reports estimates for this Markov switching model. The estimates look reasonable and significant. Notice that the heteroscedasticity-consistent standard error estimates are larger than the conventional standard error estimates reported in Hamilton (1989).

The Markov switching model reduces to the AR(4) under the constraint

$$H_0\colon \mu_d = 0.$$

Since the models are nested, two conventional statistics to test the null hypothesis would include the likelihood ratio statistic and the t-statistic for μ_d. These can be derived from Table II. These test statistics, however, do not have standard null distributions. Two reasons are paramount. First, under the null hypothesis the transition probabilities p and q are not identified. As mentioned in the introduction, this means that the large sample likelihood surface is flat (under the null) with respect to these parameters. The asymptotic likelihood has no unique maximum and is not locally quadratic. Second, the scores with respect to μ_d, p, and

Table II. Maximum-likelihood estimates of Hamilton Markov switching model, US real GNP, 1952:2 to 1984:4

Parameter	Estimate	Standard error
μ	−0·359	0·465
μ_d	1·522	0·464
φ_1	0·013	0·164
φ_2	−0·058	0·219
φ_3	−0·247	0·148
φ_4	−0·213	0·136
σ	0·769	0·094
p	0·904	0·033
q	0·755	0·101
Log-likelihood	−181·263	
LM stability test	1·364	(insignificant at 20 per cent level)

q are identically zero when evaluated at the null hypothesis. A bit of experimentation indicated that higher-order derivatives were also zero. Either of the conditions is sufficient to render standard distributional theory inapplicable.

Alternative testing procedures need to be considered. Davies (1977, 1987) and Hansen (1991) have developed testing procedures which account for the presence of unidentified nuisance parameters, but their tests and theories do not allow for identically zero scores. There is also a large literature for testing non-nested hypotheses, but these methods are not applicable since in the present context the models are nested.

One might also consider Monte-Carlo simulation. In principle this should be fairly straightforward, as it simply involves repeated fitting of Markov switching models to simulated autoregressive processes, and tabulating the resulting distributions of the likelihood ratio and/or t-statistic. Such methods have been used by Lam (1990) and Cecchetti, Lam and Mark (1990). In the evaluation of Markov switching models, however, Monte-Carlo results should be interpreted very cautiously. First, since no asymptotic theory is available for these test statistics, it is not clear whether or not the finite sample distribution will be approximately invariant to nuisance parameters (such as the density of the underlying innovations). If the large-sample distribution is not invariant to such nuisances (and this is in fact suggested by the distributional theory of Hansen (1991) for the simpler case of unidentified nuisance parameters), then it is not clear in which sense the Monte-Carlo simulations can be viewed as approximations to the true finite sample distribution.

Second, obtaining an actual Monte-Carlo draw from the required null distribution is extremely difficult. The likelihood function is severely ill-behaved, usually with numerous local optima. This problem is particularly acute when the data have been generated under the null hypothesis. The typical method of Monte-Carlo analysis in this context is to generate the data according to the null model, and fit the Markov switching model using a nonlinear maximization routine. But such routines require starting values, and their choice can have a dramatic influence upon which local maxima is found. It is quite possible (in fact, very likely) that the global maximum will be left undetected. This means that the 'likelihood ratio' statistic generated by the Monte-Carlo study will be an *underestimate* of the true likelihood ratio.[4] In

[4] Unless the Monte-Carlo analysis explicitly uses a large set of starting values, which of course increases the computational requirements.

other words, the tabulated 'distribution' will actually be a *lower bound* for the true distribution. This problem would be especially severe if the nonlinear routines use starting values which are close to the null hypothesis (since the null is often a local maxima). As a result, p-values generated by Monte-Carlo simulation (as in the aforementioned papers) can only be viewed as *liberal*: the reported p-values are *lower bounds* for the true p-values (that is, they overstate the statistical significance of the fitted model). This issue has been raised before (Hamilton, 1990), but seems to have been ignored in most Monte-Carlo studies.

In contrast, the testing procedure of section 2 produces *conservative* p-values. This paper will pursue this approach. In the notation of section 2, $\beta = \mu_d$, $\gamma = (p, q)$, and $\theta = (\mu, \sigma^2, \varphi_1, \varphi_2, \varphi_3, \varphi_4)$. The test requires computing the constrained estimates of θ for each combination of $\alpha = (\mu_d, p, q)$ using some grid of values. Throughout this section, I used for μ_d the range $[0{\cdot}1, 2]$ in steps of $0{\cdot}1$ (20 gridpoints). For the transition probabilities p and q I used three different grids in the empirical applications:

Grid 1: $0{\cdot}20$ to $0{\cdot}80$ in steps of $0{\cdot}20$ (four gridpoints);
Grid 2: $0{\cdot}15$ to $0{\cdot}90$ in steps of $0{\cdot}15$ (six gridpoints);
Grid 3: $0{\cdot}12$ to $0{\cdot}89$ in steps of $0{\cdot}11$ (eight gridpoints).

Grids 1 to 3 imply partitions of the space for (μ_d, p, q) into 320, 720 and 1280 gridpoints, respectively. Calculation of the results for each of these three choices should help to reveal the sensitivity of the test to the choice of grid. In order to achieve some efficiency in this estimation, for each value of p and q, I started with $\mu_d = 0{\cdot}1$, and used for starting values the null estimates (which correspond to $\mu_d = 0$). After convergence was obtained, I moved on to $\mu_d = 0{\cdot}2$, and used for starting values the final values from the previous optimization, and so on. This keeps the computation time down to a reasonable degree, and seems to produce the correct results. Some experimentation suggested that for (μ_d, p, q) fixed, the likelihood is well-behaved, with a single mode.

Table III presents the standardized LR statistics for this model. The test statistics, their associated p-values, and the CPU requirements are reported for the three alternative grids. The asymptotic p-values are calculated according to the method of section 3.2 using 1000 Monte-Carlo samples. The standardized LR statistics are approximately the same value, $1{\cdot}55$, for Grid 2 and Grid 3.[5] If a standard normal theory were applicable, this statistic would not reject the null hypothesis at the 5 per cent level based on the one-sided critical values, but it

Table III. Standardized LR statistics for Hamilton model

	LR_n^*	p-value	CPU hours[a,6]
Grid 1	1·24	0·77	6
Grid 2	1·56	0·68	18
Grid 3	1·55	0·72	32

[a] All computations were performed on a 486/33 in GAUSS 386. Reported computation times are in some cases estimates since the computer was occasionally running more than one program simultaneously.

[5] The standardized LR function was maximized over Grid 3 at $\mu_d = 0{\cdot}8$, $p = 0{\cdot}89$, and $q = 0{\cdot}56$.

[6] The calculations used an old version of the GAUSS386 OPTMUM application module. Thomas Goodwin has recently informed me that if the version 3·0 OPTMUM module is used instead, the CPU requirements are reduced approximately by one-half.

would be close. The standard normal theory, however, is *not* applicable, since the standardized likelihood surface has been maximized over a large number of points! This is reflected in the calculated p-values. The smallest is 0·68, which is far from significant. The approximate 5 per cent critical value is 3·0. The result is unambiguous. The AR(4) model is not rejected, and the standardized LR test fails to find any evidence in favour of the Markov switching model.

4.2. Finite Sample Distributions

In order to intelligently interpret the empirical results, it is informative to examine some simulation results to assess the actual size and power of the test in the present context. Ideally, we would like to calculate the size and power for Hamilton's actual model. Unfortunately, the computation requirements are enormous, so it seems more sensible to present results for a simplified model in which there is no autoregressive component.

To calculate rejection frequencies under the null hypothesis, I generated 50 samples of length 131, consisting of i.i.d. normal observations. Grid 2 was used to calculate the test statistic and its associated p-value. The test was deemed to reject at the 20, 10, or 5 per cent level if its associated p-value was smaller than 0·20, 0·10, or 0·05, respectively. Rejection frequencies are reported in the first line of Table IV. They are very close to those expected if the actual size of the tests were their nominal values. This is very good news, for it suggests that the fact that the asymptotic distribution is only a bound may not be very important in practice.

The most important issue, however, is effective power (rejection frequency under the alternative). Unless the test rejects the null with high probability under the alternative, the test will not provide much discriminatory power. To assess power an alternative model needs to be chosen. I used the Hamilton's point estimates for the Markov switching model, setting the autoregressive parameters equal to zero, and used Grid 2 to calculate the test statistics. Rejection frequencies are reported (again, 50 replications were made) in the second line of Table IV. The test has excellent power. A test of nominal size 5 per cent rejects the null at an 82 per cent rate under this alternative. Since the actual model has an autoregressive component, however, we cannot draw a direct conclusion concerning power in this context.

Table IV. Monte-Carlo size and power, no autoregressive component (percentages)

Nominal size	20	10	5
Size	26	10	2
Power	92	86	82

This table reports the frequency (in 50 trials) of rejections of the null hypothesis of a one-state model. The null model is i.i.d. normal innovations. The alternative model is that implied by the point estimates in Table II, with the autoregressive parameters set to zero.

4.3. Finite Sample Evidence from an Alternative Switching Model

In Hamilton's Markov switching model for GNP, the difference between states of the world is completely captured by differences in the mean of the process. When combined with an AR(4) specification, this implies that the conditional distribution of a realization depends upon the

previous four values of the Markov process. This is equivalent to a 16-state Markov process. This is the primary source of the intensive CPU requirements. In later work, Hamilton (1990, 1991a, b) emphasizes an alternative Markov switching model. Instead of having the *mean* of the process switch between states, one can have the *regression parameters* switch between the states. The obvious analogue to the mean of the process is the intercept, so an analogous specification is to have the *intercept* switch between the states. In the notation of the previous section we could write this model as:

$$\varphi(L)x_t = \mu + \mu_d s_t + e_t, \tag{12}$$

where s_t is defined as before.

(12) might seem a rather minor modification of model (11), but actually the two models have in general quite different dynamic behaviour. Examining the point estimates in Table II, however, this does not seem very important in the present context, since the autoregressive parameters are quite small, indicating that most of the dynamics have been captured by the Markov process.

The estimates from this alternative Markov switching model are given in Table V. The point estimates are quite similar to those in Table II, but the standard errors are generally smaller and the log-likelihood is higher. It appears that the modified model performs even better than Hamilton's model, although the difference is probably not statistically significant.[7]

The standardized LR statistics, reported in Table VI, were first computed for the three previously used grids on the parameters (μ_d, p, q). The statistics and their associated p-values depend somewhat upon the choice of grid, so two finer grids were also used for p and q:

Grid 4: 0·10 to 0·925 in steps of 0·075 (12 gridpoints)
Grid 5: 0·10 to 0·90 in steps of 0·05 (17 gridpoints).

Grid 5 achieves the highest value of the standardized LR statistic, at 2·23, with an associated p-value of 0·32. The associated p-values vary among the choices of grid, ranging between 0·32 and 0·46, if we exclude Grid 1 as being too coarse to be reliable. The p-values are decreasing

Table V. Maximum-likelihood estimates of Markov switching intercept model, US real GNP, 1952:2 to 1984:4

Parameter	Estimate	Standard error
μ	−0·447	0·305
μ_d	1·560	0·245
φ_1	0·112	0·105
φ_2	0·065	0·081
φ_3	−0·126	0·080
φ_4	−0·136	0·091
σ	0·789	0·066
p	0·912	0·032
q	0·669	0·143
Log-likelihood	−180·184	
Stability test	0·954	(insignificant at 20 per cent level)

[7] One could use a formal non-tested likelihood test such as that of Vuong (1989), but it does not appear to be worth the effort.

Table VI. Standardized LR statistics for Markov switching intercept model

	LR_n^*	p-Value	CPU hours
Grid 1	2·20	0·25	1·2
Grid 2	1·95	0·46	2·5
Grid 3	2·09	0·40	4·5
Grid 4	2·21	0·36	7·5
Grid 5	2·23	0·32	19·5

somewhat as we move from Grid 2 to Grid 5, but the statistics do not appear to be close to statistically significant. Again, we come to the conclusion that we are unable to reject the hypothesis of a one-state autoregression in favour of a two-state switching model.

In the above specification the conditional likelihood only depends upon the current state. As a result the computational burdens are much less demanding than those reported in section 4. This allows us to conduct a more extensive Monte-Carlo experiment for this model allowing for autoregressive components. The computational requirements are still expensive, so I focused initially upon a model which allows for an AR(1). For an evaluation of size the null data were generated according to a Gaussian AR(1). For power the data were calculated according to the point estimates from Table V, although the second to the fourth autoregressive parameters were set to zero. The results are reported in Table VII. As found in the previous experiment, the actual size of the test appears to be remarkably close to nominal values. (Note that the standard errors for the estimates of the rejection frequencies are approximately 0·05, so the exact values of these rejection frequencies should not be taken as precise.) The power of the test is also quite good, but appears to be reduced relative to the specification which did

Table VII. Monte-Carlo size and power with AR components (percentages)

Nominal size	20	10	5
AR(1) and Grid 1:			
Size	22	18	6
Power	60	36	26
AR(1) and Grid 2:			
Size	22	10	0
Power	64	50	36
AR(1) and Grid 3:			
Size	20	12	10
Power	60	48	30
AR(4) and Grid 2:			
Power	54	36	24

This table reports the frequency (in 50 trials) of rejections of the null hypothesis of a one-state model. The null model in all cases is a Gaussian AR(1). The alternative model is the model implied by the point estimates in Table II, with the second to fourth autoregressive parameters set to zero when only an AR(1) is allowed in the estimated dynamics.

not allow for an autoregressive component. For example, a 10 per cent size test using Grid 2 has a reduction of power from 86 to 50 per cent. This is not surprising, since Markov processes and autoregressive processes have similar covariance structures. It is also interesting to examine the behaviour of the tests across the different grids. It appears that Grid 1 is not sufficiently fine to achieve good power properties, but the refinement from Grid 2 to Grid 3 does not improve power.

Since the test suffered a loss in power in moving from no autoregressive component to an AR(1), one might wonder if power would continue to erode if a higher autoregressive order were allowed. Since this increases the computational requirements even further, I simplified the study by confining attention to an analysis of power under Grid 2. The other models had shown no indication of size distortion, and no gain in refining the grid, so I believe that an analysis of the other cases would not be worth the expense. The data were generated according to the point estimates from Table V. These results are reported in the last row of Table VII. They show the test retaining significant, but somewhat diminished, power. For a 10 per cent size test the probability of rejection under the alternative model is 36 per cent. For a 20 per cent size test the probability of rejection is 54 per cent. The power is not excellent. Using the analysis of Andrews (1989) we are able to conclude, however, that the Markov switching specification is 'unlikely'. This is because the observed p-value of 0·32 is above 0·20, an event which occurs less than half of the time under the alternative. This is certainly not conclusive evidence, but reinforces our view that the Markov specification has not been shown to be statistically significant.

5. A SIMPLE SWITCHING MODEL OF GNP

As was noted in the first paragraph of section 4.3, Hamilton's recent modification of the Markov switching model allows for any of the regression parameters to switch between the states. Model (11) allows only the intercept to switch. It seems odd (or at least unnecessary) to impose this restriction *a priori*. DeLong and Summers (1988), for example, argue that during the Great Depression, shocks to GNP were more persistent. They suggest that shifting autoregressive parameters can capture this phenomenon.

We can easily relax the assumption that only the intercept varies between states. The first two columns in Table VIII (under 'Unconstrained') report the estimates from a fully unrestricted model, in which the intercept, slope parameters, and error variance are all allowed to shift between the two states. All of the parameters with the 'd' subscript denote the *difference* in coefficients between states.

These estimates give a very different picture from the Markov model of Table V. The restrictions implied by the shifting intercept model are rejected by a Wald test (see Table IX) at the 1 per cent level. There appears to be a second shifting parameter, the second AR lag. In the switching-intercept model of Table V, all of the autoregressive parameters were small in magnitude, but in the unconstrained switching-parameter model, the first two AR parameters are relatively large, with the third and fourth lags relatively small.

The transition probabilities (p and q) also tell a different story. The unconstrained estimates are much smaller, and sum to 1·026. This suggests that the constraint $q = 1 - p$ should be satisfied. This constraint is of importance for two reasons. First, it eliminates one unidentified nuisance parameter, making the testing problem better-behaved. Second, the model has a different interpretation. $p + q = 1$ implies that there is *no persistence* in the Markov process, for the probability that s_t takes on one or zero is independent of the previous state. This is a *simple switching* model, rather than a *Markov switching* model, because the states arrive

Table VIII. Maximum-likelihood estimates of Markov switching parameters model, real GNP, 1952:2 to 1984:4

	Unconstrained		Constrained	
Parameter	Estimate	Std. Error	Estimate	Std. Error
μ	−0·690	0·400	−0·756	0·169
μ_d	1·815	0·362	1·871	0·158
φ_1	0·321	0·211	0·321	0·079
φ_2	0·510	0·228	0·461	0·115
φ_3	−0·078	0·121		
φ_4	−0·022	0·148		
φ_{1d}	−0·005	0·215		
φ_{2d}	−0·596	0·153	−0·582	0·133
φ_{3d}	0·006	0·189		
φ_{4d}	0·010	0·356		
σ	0·657	0·121	0·650	0·078
σ_d	0·013	0·255		
p	0·638	0·471	0·619	0·072
q	0·388	0·299		
Likelihood		−174·388		176·990
Stability		1·463		0·624

Table IX. Wald tests

Test	Statistic	Degrees of freedom	p-Value
Unconstrained vs. switching intercept	15·8	5	0·007
Unconstrained model vs. simple switching parameters	1·07	7	0·994

independently over time. It should also be appropriate to call this model a *mixture* model, since one can think of the parameters as random variables coming from a mixture distribution.

Jointly testing the seven restrictions ($\varphi_3 = \varphi_4 = \varphi_{1d} = \varphi_{3d} = \varphi_{4d} = \sigma_d = 0$ and $p + q = 1$) yields a Wald statistic of only 1·07 (see Table IX), which, as expected from an examination of Table VIII, is insignificant using a conventional chi-squared distributional approximation.[8] Since these restrictions appear to be supported by the data, it makes sense to re-estimate the model while imposing these constraints. Estimates of the restricted model are reported in the last two columns of Table VIII (under 'Constrained'). The parameter estimates are very close to the unconstrained estimates, but the estimated standard errors are much smaller.

Is this switching model statistically significant? That is, will a valid test statistic reject the null of an autoregressive model in its favour? This question deserves careful attention and evaluation. We should be careful not to make simple, yet egregious, mistakes. For example, one could look at Table VIII, and make the following claim: 'There are two parameters which vary between the states: μ and φ_2. As long as one of the two parameters is state-dependent, the transition probabilities are identified, and the scores are not identically zero under the null, hence the asymptotic distribution of the other parameter will be normal. Since the t-statistics

[8] The chi-squared distributional approximation for this and other Wald test statistics discussed in this section are valid under the tenuous assumption that the model is identified. If the model is not identified, however, the Wald statistics for selecting among competing specifications have unknown distributions.

of the two switching parameters are 11·8 and −4·4, we can confidently conclude that the switching model is statistically significant'. This argument, which appears in the existing literature (for example, Engle and Hamilton, 1990, and Hamilton, 1991a, b), sounds convincing, especially to researchers predisposed towards switching models. The argument confuses, however, the relevant testing problem. The relevant null hypothesis is a one-state model (an autoregression). Under this null, the transition probabilities are not identified and some scores are identically zero. There is no short-cut solution.

Still, even if we agree that the relevant null is a one-state model, what is the appropriate alternative? Since the constrained model on the right-side of Table VIII seems to capture all of the information in the unconstrained model, one might argue that this is the appropriate alternative. Since the parameter space is smaller, and there is only one unidentified parameter, the test would have more power against this alternative than if the unrestricted model played this role. This argument would make sense if we really could rely on the second Wald test reported in Table IX, which tests the unconstrained versus constrained models. As mentioned above, this test statistic has an asymptotic chi-square distribution when the model is identified, but the distribution is unknown when the model is not identified. The use of the constrained model for the alternative model for a hypothesis test also induces a classic pretest bias.

On the other hand, we have few alternatives. Testing the one-state null against the general unconstrained model would be hopelessly costly given my present computer resources. Computational time increases multiplicatively in the number of parameters allowed to switch between states. For example, if Grid 2 is used for p and q, and a 20-point grid for the switching parameters is used, my current GAUSS program would take about 1598 years to compute the test statistic! If a 10-point grid for the switching parameters is used, the computations would take 'only' 51 years. Of course, a more clever program could probably be written which could do the computations in less time, but the improvements would have to be enormous in order to make such computations feasible.

I therefore present a set of computationally feasible results, for two alternative hypotheses. In all cases the null is taken to be an AR(4), in order to be compatible with the earlier results and sample sizes.[9] The results differ on the grids used on p and q, and on the alternative model estimated. These alternatives can be described by the constraints they impose upon the general model.

Null model:	$\mu_d = \varphi_{1d} = \varphi_{2d} = \varphi_{3d} = \varphi_{4d} = \sigma_d = 0$
Model A:	$\varphi_{1d} = \varphi_{3d} = \varphi_{4d} = \sigma_d = 0, \quad q = 1 - p$
Model B:	$\varphi_{1d} = \varphi_{3d} = \varphi_{4d} = \sigma_d = 0$

Model A is similar to the constrained model of Table VIII, except that φ_3 and φ_4 are not constrained to equal zero. Model B is the same as Model A, except that the transition probabilities are not constrained to such unity. When testing the null against Model B, there are two unidentified nuisance parameters as in the earlier testing situations.

Computation time is roughly proportional to the total number of gridpoints, which increases exponentially in the number of gridpoints per parameter. Therefore, for these calculations I used for a grid for μ_d the range $[0{\cdot}2, 2]$ in steps of 0·2, and for φ_{2d} I used the range $[-1, 0{\cdot}8]$ in steps of 0·2 (each has 10 gridpoints). For the transition probabilities p and q I used Grids 1 through 4.

[9] In an earlier version of the paper I took the null to be an AR(2), since this is the linear restriction of the constrained model. The results were not meaningfully different from those presented here.

Table X. Standardized likelihood ratio statistics for switching parameters model

	LR_n^*	p-Value	CPU hours
Model A (μ, φ_2) vary between states, $q = 1 - p$			
Grid 1	3·50	0·02	1·6
Grid 2	3·53	0·01	2·4
Grid 3	3·61	0·01	3·4
Grid 4	3·59	0·01	5·2
Model B (μ, φ_2) vary between states, q unconstrained			
Grid 1	3·50	0·02	6
Grid 2	3·55	0·03	15
Grid 3	3·61	0·03	26
Grid 4	3·61	0·02	63

The results are reported in Table X. For both alternative models the standardized LR statistic and its associate p-value are not very sensitive to the choice of grid. In particular, Grids 3 and 4 give essentially the same answer. When testing the one-state model against Model A, the one-state model rejects at about the 1 per cent level, with a standardized LR statistic of 3·6. If this test is taken to be the appropriate test statistic, we can strongly reject the null of a linear one-state model in favour of this two-state switching model.

As mentioned above, this test is somewhat suspect, however, since Model A has been selected as a simplification of the general unconstrained model reported in Table VIII. The most important simplification, I believe, is the constraint that $p + q = 1$, since this eliminates one of the unidentified nuisance parameters. We can assess the sensitivity of our results to this problem by considering the test of the one state model against Model B. The standardized LR statistic has nearly the same value as for Model A, with only a slight increase in the associate p-value, to 0·02. This demonstrates that the computed p-value is fairly robust to the alternative model considered. We conclude that there is strong evidence against the linear one-state model.

6. CONCLUSION

This paper has set out to develop a method of hypothesis testing for nonlinear models which does not necessarily satisfy the standard list of regularity conditions. With the growing popularity of nonlinear models, more attention should be paid to regularity conditions and their violation. Statistical tools to conduct inference when regularity conditions are violated are noticeably absent.[10] This paper proposes a new and quite different approach to the subject. Essentially, the suggestion is to view the likelihood surface as the sum of the limit function and an empirical process. Random variation in estimation is entirely due to the interplay between the limit function and the random empirical process. While all we may know about the limit of the likelihood surface is that it is maximized at the null value, we can calculate the asymptotic distribution of the likelihood empirical process from the data themselves. This enables us to bound the distribution of the maximum of the standardized likelihood ratio process, and use this maximum as a test of the null hypothesis.

[10] The one issue which has been discussed at length is estimation and testing subject to boundary conditions. See, for example, Chernoff (1954), Moran (1971), Gourieroux *et al.* (1982), Rogers (1986), and Wolak (1989).

This paper also investigates the statistical significance of Hamilton's (1989) Markov switching model for GNP. The violations of the conventional regularity conditions are strong, and I am unable to reject the hypothesis that the 'good fit' of Hamilton's model is simply due to sampling error. Instead, I estimate an alternative which is a simple switching model of GNP, which allows both the intercept and the second AR parameter to randomly shift between two values. This switching model fits the data better than an AR(4), rejecting the latter at the asymptotic 1 per cent level.

ACKNOWLEDGEMENTS

I would like to thank Fabio Canova, Adrian Pagan, Simon Potter and Changyong Rhee for helpful conversation during this project, two referees for constructive criticisms, James Hamilton for helpful comments and kindly providing his data and computer programs, and the NSF for financial support under grant SES 9022176.

REFERENCES

Andrews, D. W. K. (1989), 'Power in econometric applications', *Econometrica*, **57**, 1059–1090.

Andrews, D. W. K. (1991), 'An empirical process central limit theorem for dependent non-identically distributed random variables', *Journal of Multivariate Analysis*, **38**, 187–203.

Andrews, D. W. K., and D. Pollard (1990), 'A functional central limit theorem for strong mixing stochastic processes', Cowles Foundation Discussion Paper 951.

Cecchetti, S. G., P. S. Lam, and N. C. Mark (1990), 'Mean reversion in equilibrium asset prices', *American Economic Review*, **80**, 398–418.

Chernoff, H. (1954), On the distribution of the likelihood ratio', *Annals of Mathematical Statistics*, **25**, 573–578.

Davies, R. B. (1977), 'Hypothesis testing when a nuisance parameter is present only under the alternative', *Biometrika*, **64**, 247–254.

Davies, R. B. (1987), 'Hypothesis testing when a nuisance parameter is present only under the alternative', *Biometrika,* **74**, 33–43.

DeLong, J. B., and L. H. Summers (1988), 'On the existence and interpretation of a "unit root" in U.S. GNP', *NBER Working Paper Series*, No. 2716.

Engle, C., and J. D. Hamilton (1990), 'Long swings in the dollar: are they in the data and do markets know it?', *American Economic Review*, **80**, 689–713.

Gourieroux, C., A. Holly, and A. Monfort (1982), 'Likelihood ratio test, Wald test and Kuhn–Tucker test in linear models with inequality constraints on the regression parameters', *Econometrica*, **50**, 63–80.

Hamilton, J. D. (1989), A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica*, **57**, 357–384.

Hamilton, J. D. (1990), 'Specification testing in Markov-switching time-series models'. University of Virginia.

Hamilton, J. D. (1991a), 'Estimation, inference, and forecasting of time series subject to changes in regime', *Handbook of Statistics*, vol. 10 (forthcoming).

Hamilton, J. D. (1991b), 'State–space models', *Handbook of Econometrics*, vol. 4 (forthcoming).

Hansen, B. E., (1990), 'Lagrange multiplier tests for parameter instability in non-linear models', University of Rochester.

Hansen, B. E. (1991), 'Inference when a nuisance parameter is not identified under the null hypothesis', University of Rochester Discussion Paper 296.

Horowitz, J. L., and J. McAleer (1989), 'A simple method for testing a general parametric model against a non-tested alternative', University of Iowa.

Kemp, G. C. R. (1991), 'On Wald tests for globally and locally quadratic restrictions', *Journal of Econometrics*, **50**, 257–272.

Lam, P. S. (1990), 'The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series', *Journal of Monetary Economics*, **26**, 409–432.

Lee, L. F., and A. Chesher (1986), 'Specification testing when score test statistics are identically zero', *Journal of Econometrics*, **31**, 121–149.

Moran, P. A. P. (1971), 'Maximum-likelihood estimation in non-standard conditions', *Proceedings of the Cambridge Philosophical Society*, **70**, 441–450.

Nyblom, J. (1989), 'Testing for the constancy of parameters over time', *Journal of the American Statistical Association*, **84**, 223–230.

Rogers, A. J. (1986), 'Modified Lagrange multiplier tests for problems with one-sided alternatives', *Journal of Econometrics*, **31**, 341–361.

Vuong, Q. H. (1989), 'Likelihood ratio tests for model selection and non-tested hypotheses', *Econometrica*, **57**, 307–333.

White, H. (1980), 'A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica*, **48**, 817–838.

Wolka, F. A. (1989), 'Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models', *Econometric Theory*, **5**, 1–35.

JOURNAL OF APPLIED ECONOMETRICS, VOL. 11, 195–198 (1996)

ERRATUM: THE LIKELIHOOD RATIO TEST UNDER NONSTANDARD CONDITIONS: TESTING THE MARKOV SWITCHING MODEL OF GNP

BRUCE E. HANSEN
Department of Economics, Boston College, Chestnut Hill, MA 02167-3806, USA

There was an error in Hansen (1992). I am very grateful to James Hamilton for pointing out the error.

Equations (2) and (3) in the original read

$$\frac{1}{\sqrt{n}} Q_n(\alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q_i(\alpha) \Rightarrow Q(\alpha) \tag{2}$$

where $Q(\alpha)$ is a mean zero Gaussian process with covariance function

$$K(\alpha_1, \alpha_2) = E(q_i(\alpha_1) q_i(\alpha_2)). \tag{3}$$

While equation (2) is correct, (3) is not. Instead, the correct expression is

$$K(\alpha_1, \alpha_2) = \sum_{k=-\infty}^{\infty} E(q_i(\alpha_1) q_{i+k}(\alpha_2)) \tag{3'}$$

The reason is that the likelihood components $q_i(\alpha)$ will be serially correlated for some values of α. This will be the case even when the original data are iid, since the likelihood $q_i(\alpha)$ is a function of all data up to time i. It should be noted that this problem does not apply to the testing methods of Hansen (1994), which involve application of empirical process theory to specific likelihood scores which are serially uncorrelated.

This error implies that the method of calculating the asymptotic distribution in Section 3.2 is incorrect. Instead, set $\hat{q}_i(\alpha) = q_i(\alpha, \hat{\theta}(\alpha))$ and

$$\hat{K}_n(\alpha_1, \alpha_2) = \sum_{i=1}^{n} \hat{q}_i(\alpha_1) \hat{q}_i(\alpha_2) + \sum_{k=1}^{M} w_{kM} \left[\sum_{1 \leq i \leq n-k} \hat{q}_i(\alpha_1) \hat{q}_{i+k}(\alpha_2) + \sum_{1+k \leq i \leq n} \hat{q}_i(\alpha_1) \hat{q}_{i-k}(\alpha_2) \right] \tag{3'}$$

where $w_{kM} = 1 - |k|/(M+1)$ is the Bartlett kernel and M is a bandwidth number (selected to grow to infinity slowly with sample size). Then a consistent estimate of

$$K^*(\alpha_1, \alpha_2) = \frac{K(\alpha_1, \alpha_2)}{V(\alpha_1)^{1/2} V(\alpha_2)^{1/2}}$$

CCC 0883–7252/96/020195–4

Received June 1994
Revised May 1995

is given by

$$K_n^*(\alpha_1, \alpha_2) = \frac{\hat{K}_n(\alpha_1, \alpha_2)}{V_n(\alpha_1)^{1/2} V(\alpha_2)^{1/2}}$$

Sample draws from this process can be obtained by constructing

$$\widetilde{LR}^*(\alpha) = \frac{\sum_{k=0}^{M} \sum_{i=1}^{n} q_i(\alpha, \hat{\theta}(\alpha)) u_{i+k}}{\sqrt{1+M}\, V_n(\alpha)^{1/2}}$$

where $\{u_i\}_{i=1}^{n+M}$ is a sample of random $N(0,1)$ variables. The reader may verify that conditional on the data, $\widetilde{LR}^*(\alpha)$ is a mean zero Gaussian process with exact covariance function $K_n^*(\alpha_1, \alpha_2)$, and the latter is an asymptotic approximation to $K^*(\alpha_1, \alpha_2)$.

The theory does not give any particular guidance for choice of M. It therefore seems prudent to calculate the tests for several choices to assess sensitivity.

The original paper reported estimates of $K_n^*(\alpha_1, \alpha_2)$ and $\widetilde{LR}^*(\alpha)$ effectively with $M = 0$. Thus all test statistics and Monte Carlo evidence were presented with $M = 0$. All the numerical work was recalculated for $M = 1, \ldots, 4$. Other than the change discussed above, the methods were essentially identical to those outlined in Hansen (1992). The corrected results are presented in the following tables. It it interesting to note that the results are not very sensitive to M. None of the conclusions drawn are affected. The category 'CPU hours' referred to the time required for the programs to run on a 486/66 computer.

A GAUSS program which produces the empirical results reported here is available on request from the author.

Table III. Standardized LR statistics for Hamilton model

		p-value					
	LR_n^*	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	CPU hours
Grid 1	1·24	0·77	0·75	0·75	0·76	0·73	1·2
Grid 2	1·56	0·73	0·68	0·67	0·66	0·62	2·5
Grid 3	1·55	0·74	0·70	0·69	0·69	0·65	4·6

Table IV. Monte Carlo size and power, no autoregressive component (percentages)

		Null			Alternative		
Nominal size:	M	20	10	5	20	10	5
	0	12	4	0	86	80	74
	1	14	8	0	86	80	74
	2	16	8	2	86	76	74
	3	14	8	2	86	76	74
	4	14	6	2	86	76	74

Table VI. Standardized LR statistics for Markov switching intercept model

		p-value					
	LR_n^*	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	CPU hours
Grid 1	2·20	0·29	0·26	0·30	0·27	0·28	0·3
Grid 2	1·95	0·49	0·51	0·50	0·46	0·48	0·6
Grid 3	2·09	0·44	0·42	0·43	0·41	0·43	1·1
Grid 4	2·21	0·41	0·38	0·38	0·37	0·38	2·6
Grid 5	2·23	0·39	0·38	0·36	0·34	0·34	5·2

Table VII. Monte Carlo size and power with AR components (percentages)

		Null			Alternative		
Nominal size:	M	20	10	5	20	10	5
AR(1)	0	14	10	6	52	40	30
	1	14	10	6	52	40	30
	2	14	10	6	52	38	28
	3	16	10	4	52	38	28
	4	16	10	4	52	36	26
AR(4)	0	26	18	14	44	36	24
	1	30	20	16	44	40	24
	2	28	20	16	44	40	22
	3	28	22	16	44	38	24
	4	30	22	16	44	33	20

Table X. Standardized LR statistics for switching parameters model

		p-value					
	LR_n^*	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	CPU hours
Model A: (μ, ϕ_2) vary between states, $q = 1 - p$							
Grid 1	3·50	0·01	0·01	0·01	0·02	0·01	0·4
Grid 2	3·53	0·01	0·01	0·01	0·02	0·02	0·6
Grid 3	3·61	0·01	0·02	0·01	0·01	0·01	0·8
Grid 4	3·59	0·01	0·01	0·01	0·01	0·01	1·2
Model B: (μ, ϕ_2) vary between states, q unconstrained							
Grid 1	3·50	0·02	0·02	0·02	0·02	0·03	1·8
Grid 2	3·55	0·03	0·03	0·03	0·03	0·04	4·0
Grid 3	3·61	0·02	0·03	0·03	0·04	0·03	7·1
Grid 4	3·61	0·04	0·04	0·02	0·03	0·04	15·9

ACKNOWLEDGEMENTS

I thank two referees for helpful comments. Financial support from the National Science Foundation and the Sloan Foundation is gratefully acknowledged.

REFERENCES

Hansen, B. E. (1992), 'The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP', *Journal of Applied Econometrics*, **7**, S61–82.

Hansen, B. E. (1994), 'Inference when a nuisance parameter is not identified under the null hypothesis', *Econometrica*, forthcoming.