Econometric Theory, 9, 1993, 377–401. Printed in the United States of America.

DISTRIBUTION OF THE ML ESTIMATOR OF AN MA(1) AND A LOCAL LEVEL MODEL

NEIL SHEPHARD
Nuffield College, Oxford

Although considerable attention has recently been paid to the behavior of the maximum likelihood estimator of simple moving average models, little progress has been made in finding a good approximation to its distribution in cases where the process is close to being noninvertible. In this paper a method is produced that gives an excellent approximation to the distribution function, even in the case where the process is strictly noninvertible. Also studied is the related problem of the distribution of the maximum likelihood estimator of the signal-to-noise ratio in the local level model.

1. INTRODUCTION

The first-order moving average model, written as MA(1), is given by

$$y_t = \epsilon_t + \theta\epsilon_{t-1}, \qquad \epsilon_t \sim \text{NID}(0, \sigma^2), \qquad (t = 1, \ldots, T). \tag{1.1}$$

The process is said to be invertible if and only if $|\theta| < 1$. The maximum likelihood (ML) estimator of an invertible moving average is asymptotically normally distributed. On the other hand, if $|\theta| = 1$, then the process is said to be strictly noninvertible and normality no longer holds. Noninvertibility is of great practical importance for it implies that the time series has been over differenced and so shows that the order of integration has been overestimated. This paper establishes an approximate sampling theory for the ML estimator which will be accurate even in the strictly noninvertible case.

The search for a theory to cover the noninvertible case has created a considerable body of work following the early simulation experiments of Kang [25]. It was Kang who first discovered that a variety of commonly used estimators of the MA(1) process behaved strangely when $|\theta|$ was close to 1. In particular, she reported that many of her replications had estimates which seemed to be exactly on the boundary of noninvertibility. This point was reinforced by a battery of simulation studies conducted by Cooper and Thomp-

I am grateful to Dr. Steve Satchell for several discussions on this topic and Professor Peter Robinson for pointing out references [22] and [31]. The comments of the referees and Professor P.C.B. Phillips on a previous version of this paper were also very helpful. I am responsible for any errors that may remain in this paper.

 0266-4666/93 $5.00 + .00

son [5], Harvey [17, pp. 136–139], Dent and Min [12], Ansley and Newbold [3], and Davidson [8,9].

The work of Ansley and Newbold [3] is perhaps the most revealing of these papers. It compared exact ML, exact least squares, and conditional least-squares estimators for a variety of models, reporting mainly on bias, mean square error, and predictive ability. Ansley and Newbold [3, p. 181] found that ML estimators were generally more reliable than the other estimators they had studied. However, the ML estimator of the θ parameter of the MA(1) process was shown to have a large probability of occurring near the boundary of noninvertibility. Of the 10,000 replications they conducted, when the true value of θ was -0.9 and $T = 50$, 3,278 of the estimates of θ were found to be between -0.99 and -1. The authors argued that none of these estimates were exactly equal to -1 but were nearly this value since "it can be shown that $\hat{\theta}$ will take the value -1 with probability zero," (see p. 172). This assertion was shown to be false by Cryer and Ledolter [6], who derived the exact sampling distribution for the ML estimator when $T = 2$. This distribution was demonstrated to be discontinuous at -1 and 1, but continuous between these two points. Cryer and Ledolter went on to compute the exact probability of observing the estimator being on the boundary of noninvertibility for any value of T.

The first analytic work on estimated noninvertibility, carried out by Sargan and Bhargava [32], proved that the ML estimator of a noninvertible MA(1) is T-consistent, sharing the same speed of convergence as the least-squares estimator of a unit root autoregression; see White [37]. Sargan and Bhargava [32] also showed that the limiting probability of the ML estimator being exactly -1, if the true value of θ was -1, was 0.6575 (Tanaka and Satchell [35] corrected this result to 0.65744). Similar results were obtained for the case of a regression model with MA(1) errors. Pesaran [28] produced a slightly simpler analysis of the above setup, but his work did not contribute anything essentially new.

Anderson and Takemura's [2] elegant analysis of the estimation of noninvertible processes produced two advances. The first was a proof that the probability that a noninvertible process is estimated, when the true process is invertible, is $o(T^{-n})$ where n is any positive integer. The second was the setting down of a framework for the analysis of estimated noninvertibility in moving averages of a general order.

More recently, Tanaka and Satchell [35] attempted to analytically approximate the distribution of the estimator of θ when the true process is noninvertible. Although they failed to give a viable approximation, their approach was very interesting and will be referenced frequently in what follows. Pötscher [29] looked at the consistency of ML estimation under various types of misspecification.

In this paper the distribution of the ML estimator of the MA(1) process is approximated. This approximation will prove to be accurate, even when

the true process is noninvertible and the sample size is small. The technique is first developed on a constrained ARIMA(0,1,1) model, the local level model, as it is mathematically more tractable than the unconstrained MA(1). The local level model is defined as

$$y_t = \mu_t + \epsilon_t, \qquad \epsilon_t \sim \text{NID}(0, \sigma^2), \qquad (t = 1, \ldots, T+1) \tag{1.2a}$$

$$\mu_t = \mu_{t-1} + \eta_t, \qquad \eta_t \sim \text{NID}(0, q\sigma^2), \tag{1.2b}$$

$$\mu_0 | Y_0 \sim N(\mu, \kappa). \tag{1.2c}$$

where $\{\epsilon_t\}$ and $\{\eta_s\}$ are independent for all values of t and s. Further, $\mu_0 | Y_0$, where Y_0 denotes the information set available at time 0, is assumed to be independent of all the noise terms in the model. Throughout this paper κ is forced to go to infinity, so the local level model is initialized by using a diffuse prior, although exactly the same results would be achieved if κ had been taken to be zero and a marginal or restricted likelihood had been constructed by regarding μ as a nuisance parameter; see McCullagh and Nelder [26, Chapter 9], Kalbfleisch and Sprott [24], Harville [21], or Robinson [30]. Very different results would occur, however, if instead of the marginal, a profile or concentrated likelihood had been used to remove μ; see Shephard and Harvey [34] in this context and Shephard [33] in the more general case where there are also regressors present in (1.2a).

The local level model is of considerable interest in its own right. Muth [27] showed that it provides the rationale for the use of the exponentially weighted moving average scheme. It is exploited as the basis of West and Harrison's [36] dynamic linear models and Harvey's [20] structural time series models. Examples of the use of these types of models on economic data include Harvey, Henry, Peters, and Wren-Lewis [18] in their work on the employment-output equation and Harvey and Stock [19] on the income-consumption relationship.

In this model σ^2 is a scale parameter, while q is called the signal-to-noise ratio. It is q that controls the time series properties of the local level model. The reduced form of a local level model is an MA(1) model with q related to θ by $q = -(1+\theta)^2/\theta$. As q is non-negative, θ must be constrained to be nonpositive.

An approximation to the distribution function of the estimator of q will be developed. It turns out that this distribution function is quite complicated for there is a nonzero probability that q is estimated to be zero (Shephard and Harvey [34]). When the true value of q is zero, the differenced process is noninvertible and so the usual asymptotic theory for the estimator breaks down. The proposed method for calculating the distribution function is, however, still valid, allowing the tabulation of the distribution function in this case.

The rest of this paper is structured in the following way. Some straightforward results for the ML estimator of the local level model are derived in

Section 2. Section 3 introduces the key tool for deriving the approximation to the distribution function of the ML estimator of moving averages. In Section 4 this technique is applied to the problem of the local level model. It is shown that the difference between the true distribution function and the suggested approximation is minor for very small sample sizes. Section 5 includes plots of the distribution function of the ML estimator of the local level model. In Section 6 the unconstrained MA(1) problem is tackled. A summary of the results of the paper is set out in Section 7. Except where stated, the proofs are presented in the Appendix.

2. THE LOCAL LEVEL MODEL

If the sequence $\{z_t\}$ denotes the first difference of the local level model and $z = (z_2, \ldots, z_{T+1})'$, then the log-likelihood for $y = (y_1, \ldots, y_{T+1})'$ is given by

$$\lim_{\kappa \to \infty} (\log L(\sigma^2, q; y) + \tfrac{1}{2}\log \kappa) = \log L(\sigma^2, q; z). \tag{2.1}$$

B will be written to denote a $(T \times T)$ matrix with the band structure

$$B = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & & 1 & 0 \end{bmatrix}, \tag{2.2}$$

to enable the log-likelihood for the local level model to be expressed as

$$\log L(\sigma^2, q; z) = \text{const} - \frac{T}{2}\log \sigma^2 - \frac{1}{2}\log|(2+q)I - B| - \frac{z'((2+q)I - B)^{-1}z}{2\sigma^2}. \tag{2.3}$$

This can be simplified by noting that the eigenvalues of B are

$$\lambda_t = 2\cos\frac{t\pi}{T+1}, \qquad \text{for} \quad t = 1, \ldots, T,$$

([2, Section 6.5]) and by defining $\delta_t(q) = 2 + q - \lambda_t$. Then if q^* and σ^{*2} are used to denote the true values of q and σ^2, respectively, the log-likelihood can be written as

$$\log L(\sigma^2, q; z) = \text{const} - \frac{T}{2}\log\sigma^2 - \frac{1}{2}\sum_{t=1}^{T}\log\delta_t(q) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\sigma^{*2}u_t^2\frac{\delta_t(q^*)}{\delta_t(q)}, \tag{2.4}$$

where $u_t \sim \text{NID}(0,1)$. Although it is well known that scaled versions of the ML estimators of q and σ^2 are asymptotically normally distributed if their true values are strictly positive (see, for example, Harvey [20, Chapter 4]), the analytic form of their asymptotic variance is not known. It is given in Theorem 1.

THEOREM 1. *If the true values of q and σ^2 are strictly positive and their ML estimators are written as $\hat{q}$ and $\hat{\sigma}^2$, then the following holds*:

$$\sqrt{T}\begin{pmatrix}\hat{\sigma}^2 - \sigma^2\\ \hat{q} - q\end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \frac{2\sigma^4}{(q+2)-\sqrt{q(q+4)}} \times \begin{pmatrix} q+2 & -q(q+4)/\sigma^2 \\ -q(q+4)/\sigma^2 & (\sqrt{q(q+4)})^3/\sigma^4\end{pmatrix}\right), \tag{2.5}$$

as the sample size goes to infinity. ■

When the true value of q is zero, asymptotic normality fails as the reduced form of the local level model is noninvertible. In this case very little is known about its distribution. A tractable way to make progress is to construct the profile likelihood function by concentrating out σ^2 at the value

$$\hat{\sigma}^2 = \frac{z'((2+q)I - B)^{-1}z}{T}, \tag{2.6}$$

to yield the objective function

$$M(q;z) = \log L(\hat{\sigma}^2, q; z) = -\frac{1}{2}\log|(2+q)I - B| - \frac{T}{2}\log z'((2+q)I - B)^{-1}z. \tag{2.7}$$

The associated score is

$$s(q;z) = \frac{dM}{dq} = -\frac{1}{2}\operatorname{tr}(((2+q)I - B)^{-1}) + \frac{T}{2}\frac{z'((2+q)I - B)^{-2}z}{z'((2+q)I - B)^{-1}z} \tag{2.8}$$

$$= -\frac{1}{2}\sum_{t=1}^{T}\delta_t^{-1}(q) + \frac{T}{2}\frac{\sum_{t=1}^{T}u_t^2\frac{\delta_t(q^*)}{\delta_t^2(q)}}{\sum_{t=1}^{T}u_t^2\frac{\delta_t(q^*)}{\delta_t(q)}}, \tag{2.9}$$

which allows the proof of Theorem 2.

THEOREM 2. *If the true value of q is zero, then as the sample size goes to infinity the following result holds*

$$\frac{s(q;z)}{T^2} \xrightarrow{d} \frac{1}{2}\sum_{t=1}^{\infty} u_t^2 \frac{t^2\pi^2}{(\pi^2 t^2 + c)^2} - \frac{1}{2}\sum_{t=1}^{\infty} \frac{1}{(\pi^2 t^2 + c)} = r(c;u), \tag{2.10}$$

where $c = T^2 q$ and $u = (u_1, \ldots, u_T)'$. ■

Theorem 2 allows the determination of the rate of convergence of the ML estimator of q in this nonstandard case. The result is given in Theorem 3.

THEOREM 3. *The ML estimator of q is T^2-consistent if the true value of q is zero. This is taken to mean that for any $\epsilon > 0$, there exists a $c > 0$ such that*

$$\lim_{T\to\infty} \operatorname{pr}\left(\frac{s(q;z)}{T^2} \Big|_{q=c/T^2} \geq 0 \right) < \epsilon. \tag{2.11}$$
■

Super consistency results for estimators of nonstationary time series models are quite common—see, for example, White [37]. The first result of this type for noninvertible time series models was published by Sargan and Bhargava [32] who proved T-consistency of the ML estimator of an MA(1) model.

Having established the speed of convergence of the estimator, it is very natural to try to derive the distribution of $T^2\hat{q}$, where $\hat{q}$ denotes the ML estimator. This turns out to be a very difficult problem.

The most natural approach to this problem is to work with the asymptotic distribution of the score given in (2.10). To study the asymptotic distribution of $T^2\hat{q}$, it is useful to temporarily extend its permissible parameter space. Instead of working with the non-negative real line, suppose $c \in (-\pi^2, \infty)$. Then as $c \to -\pi^2$, $r(c;u) \to \infty$, while as $c \to \infty$, $r(c;u) \to -\infty$, implying there must exist at least one solution to the asymptotic likelihood equation $r(\hat{c};u) = 0$ for $\hat{c} \in (-\pi^2, \infty)$. The solution need not be unique, however, for it is quite possible that there will be multiple solutions to the asymptotic likelihood equation.

As it stands, $r(c;u)$ tells us little about the asymptotic distribution of $T^2\hat{q}$. However, the likelihood equation can be expanded by using a formal Taylor series in an attempt to achieve some tractability. This is the approach followed by Tanaka and Satchell [32] in their work on noninvertible moving averages. The corresponding result for the local level model is given in Theorem 4.

THEOREM 4. *For $\hat{c} = T^2\hat{q}$ in the range $(-\pi^2, \pi^2)$, the limiting likelihood equation can be written as*

$$0 = r(\hat{c};u) = \sum_{t=1}^{\infty} \frac{u_t^2 - 1}{t^2\pi^2} - \hat{c}\sum_{t=1}^{\infty} \frac{2u_t^2 - 1}{(t^2\pi^2)^2} + \hat{c}^2 \sum_{t=1}^{\infty} \frac{3u_t^2 - 1}{(t^2\pi^2)^3}$$
$$- \hat{c}^3 \sum_{t=1}^{\infty} \frac{4u_t^2 - 1}{(t^2\pi^2)^4} + \hat{c}^4 \sum_{t=1}^{\infty} \frac{5u_t^2 - 1}{(t^2\pi^2)^5} - \quad \blacksquare \tag{2.12}$$

Although this sum is of some interest, it has two difficulties. First, it is valid only up to π^2, which turns out to be only a very small part of the interesting range for c. Second, this representation does not allow the tabulation of the distribution function of $\hat{c}$, for it is still highly nonlinear in $\hat{c}$.

Tanaka and Satchell [32] responded to the nonlinearity of the representation of the likelihood equation by truncating the infinite sum so that only two terms remained. However, the result gives only a very poor approximation to the true distribution and so is of little practical use.

3. SAMPLING THEORY FOR MAXIMUM LIKELIHOOD ESTIMATION

A technique for approximating the distribution of extremum estimators, which offers a great deal of promise in this context, was suggested by Huber [22] in deriving the asymptotic distribution of robust location estimators. More recently, Robinson [31], Daniels [7], and Field and Hampel [14] (see also Barndorff-Nielson and Cox [4, p. 130]) have used the same idea in their work on robustness. To discuss this procedure, it will be useful to introduce some notation. Let λ denote a scalar parameter, $f(\lambda)$ some objective function, and $s(\lambda)$ the derivative of f. Further, let $\hat{\lambda}$ denote the value of λ for which f is globally maximized over the relevant range for λ. Huber used the result that if the objective function is continuously differentiable and the derivative is monotonically decreasing, then

$$\text{pr}(\hat{\lambda} \leq \lambda) = \text{pr}(s(\lambda) \leq 0). \tag{3.1}$$

The requirement of the monotonically decreasing derivative is actually not necessary for this result to hold. This idea is developed in Theorem 5.

THEOREM 5. *Let λ denote a scalar parameter, $f(\lambda)$ some nonstochastic objective function, $s(\lambda)$ the derivative of f, and $\hat{\lambda}$ the value of λ for which f is globally maximized over the range $[a,b]$. Assume*

(i) *$s(\lambda)$ is continuous for $\lambda \in [a,b]$,*
(ii) *$s(a) \neq 0$,*
(iii) *$s(b) \neq 0$,*
(iv) *there are a finite number of solutions to the equation $s(\bar{\lambda}) = 0$ and all these solutions are either maximums or minimums,*
(v) *there is exactly one point which is a local maximum in f.*

Then

$\hat{\lambda} \leq \lambda$ if and only if $s(\lambda) \leq 0$. ■

This analytic result allows the proof of Theorem 6.

THEOREM 6. *Let λ denote a scalar parameter, $f(\lambda)$ some stochastic objective function, $s(\lambda)$ the derivative of f, and $\hat{\lambda}$ the value of λ for which f is globally maximized over the range $[a,b]$. Assume:*

(i) *$s(\lambda)$ is continuous for $\lambda \in [a,b]$,*
(ii) *$\text{pr}(s(\lambda) = 0) = 0$ for all $\lambda \in [a,b]$, and all the solutions to the equation $s(\lambda) = 0$ are either maximums or minimums,*

and write U to denote the number of local maximums in f. Then for all values of λ in the range $[a,b]$

$$\text{pr}(\hat{\lambda} \leq \lambda) - \text{pr}(s(\lambda) \leq 0) = \tau(\lambda) \tag{3.2}$$

where

$$\tau(\lambda) = [\text{pr}(\hat{\lambda} \leq \lambda \mid U > 1) - \text{pr}(s(\lambda) \leq 0 \mid U > 1)]\text{pr}(U > 1). \tag{3.3}$$

■

Theorem 6 will turn out to be very powerful if $\tau(\lambda)$ is small, for then the distribution function of the ML estimator can be computed by using the approximation

$$\text{pr}(\hat{\lambda} \leq \lambda) \cong \text{pr}(s(\lambda) \leq 0), \tag{3.4}$$

for the distribution function of the score for the local level model is readily available. A drawback of this approach is that $\text{pr}(s(\lambda) \leq 0)$ is not necessarily monotonically nondecreasing in λ and hence could result in an approximate distribution function which falls as λ increases. Of course, this will not be a problem if $\tau(\lambda)$ is small across all λ.

4. THE LOCAL LEVEL MODEL REVISITED

Equation (3.4) will be used to approximate the distribution function of the ML estimator of q, with

$$\text{pr}(\hat{q} \leq q) \cong \text{pr}(s(q;z) \leq 0)$$

$$= \text{pr}\left[\sum_{t=1}^{T} u_t^2 \frac{\delta_t(q^*)}{\delta_t(q)} (\delta_t^{-1}(q) - d(q)) \leq 0\right],$$

$$\text{where} \quad d(q) = \frac{1}{T}\sum_{t=1}^{T} \delta_t^{-1}(q). \tag{4.1}$$

The probability that the score is negative can be computed exactly by using Davies' [10,11] or Farebrother's [13] procedure. Theorem 6 states that the error induced by using this approximation is

$$\tau(q;q^*) = P(\hat{q} \le q) - \text{pr}(s(q) \le 0)$$
$$= (\text{pr}(\hat{q} \le q \mid U > 1) - \text{pr}(s(q) \le 0 \mid U > 1))\,\text{pr}(U > 1). \quad \textbf{(4.2)}$$

It seems difficult to make analytic progress on evaluating any of the three expressions on the right-hand side of this equality, except in two cases. The first is where $T = 2$, in which case $\text{pr}(U > 1) = 0$; see Cryer and Ledolter [6]. The second is where $q^* > 0$ and T goes to infinity, for then Huzurbazar's [23] result that the likelihood equation will have a sole consistent root can be used to prove that $\text{pr}(U > 1) \to 0$ as $T \to \infty$. Unfortunately, Huzurbazar's result does not cover the case where q^* is exactly zero.

In response to these difficulties, a simulation experiment was designed to evaluate these three unknown probabilities. q^* was taken to be one of four values, 0, 0.01, 0.1, and 1, while T took on the values 3, 10, 20, 50, 100, and 200. For each combination, 5,000 replications were computed and the resulting log-likelihood carefully searched to find all the permissible local maximums. In each case $\text{pr}(s(q) \le 0 \mid U > 1)$, $\text{pr}(\hat{q} \le q \mid U > 1)$ and $\text{pr}(U > 1)$ were estimated. Table 1 displays the results for $\text{pr}(U > 1)$ for all values of q^* and T. This probability is an estimate of the upper bound on the possible error induced when using the approximation suggested in (4.1). It implies that this error is small for $T > 50$. However, for smaller sample sizes the errors are sometimes worrying. This is especially true when the sample size is under 20. Notice the result is exact when $T = 3$.

$\text{pr}(U > 1)$ is only one of the elements that makes up $\tau(q;q^*)$. Figure 1 displays $\text{pr}(s(q) \le 0 \mid U > 1)$ and $\text{pr}(\hat{q} \le q \mid U > 1)$ for various values of T when $q^* = 0$. Similar shapes hold for all the other values of q^* and so are not reported here. The difference $\text{pr}(s(q) \le 0 \mid U > 1;\ q^* = 0) - \text{pr}(\hat{q} \le q \mid U > 1;\ q^* = 0)$, which is explicitly plotted in Figure 2, is largest when q is small. This is caused by the occurrence of a considerable number of local maximums at zero not being global maximums. This feature does not change substantially as the sample size increases.

A similar type of result occurs in Figure 3 which gives $\tau(q;q^* = 0.1)$.

Table 1. Upper bound on the approximating error, $100 \times \text{pr}(U > 1)$

t	$q^* = 0$	$q^* = 0.01$	$q^* = 0.1$	$q^* = 1$
3	0.0	0.0	0.0	0.0
10	2.44	1.92	2.46	1.32
20	1.36	1.20	1.36	1.56
50	1.02	1.10	1.18	0.34
100	0.94	0.84	0.36	0.06
200	1.00	0.84	0.02	0.00

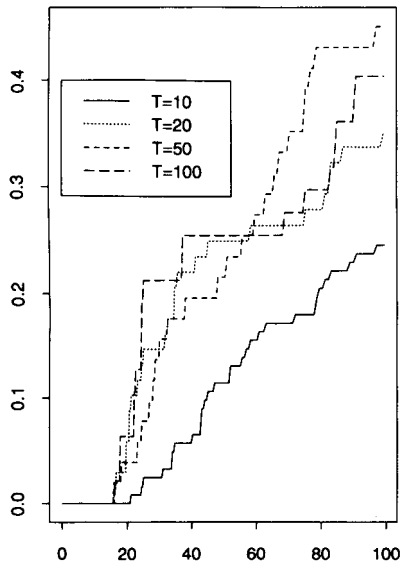

FIGURE 1a. $\text{pr}(\hat{q} \leq q | U > 1; q^* = 0)$.

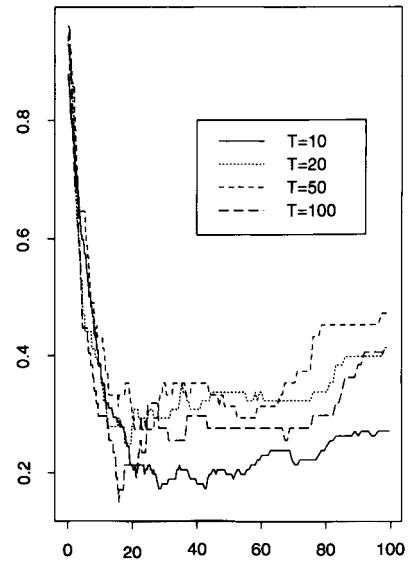

FIGURE 1b. $\text{pr}(s(q) \leq 0 | U > 1; q^* = 0)$, for various T. X-axis: $T^2 q$, Y-axis: probability.

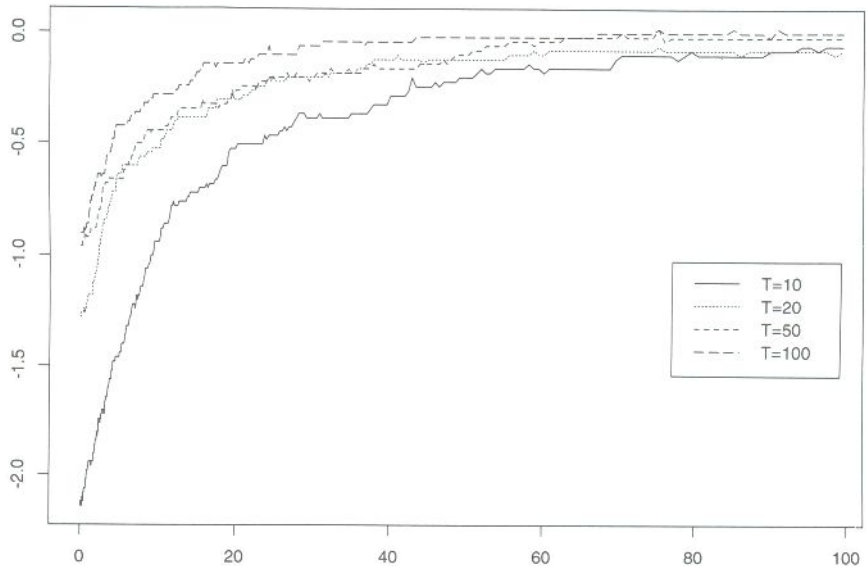

FIGURE 2. $\tau(q; q^* = 0)$ for various values of T. X-axis: T^2q, Y-axis: 100τ.

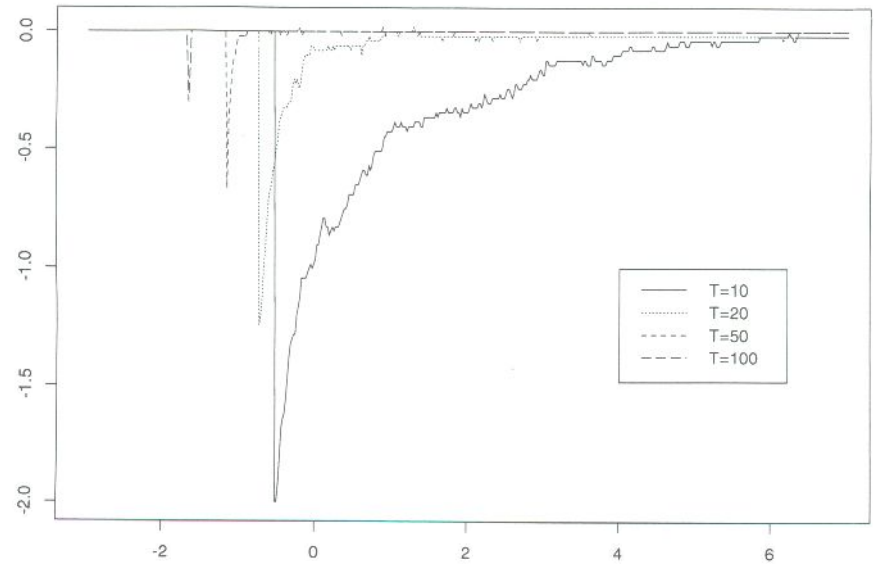

FIGURE 3. $\tau(q; q^* = 0.1)$ for various values of T. X-axis: standardized q, Y-axis: 100τ.

When T is larger than 20, τ is quite small, especially when q is large. Of course, as $T \to \infty$, $\tau(q;q^*) \to 0$ for all $q^* > 0$.

5. THE ML ESTIMATOR OF THE LOCAL LEVEL MODEL

The accuracy of the approximation suggested by Theorem 6 prompts its application. The results from the use of (4.1) are drawn in Figures 4 and 5 for $q^* = 0$ and 0.1, respectively. Each figure has five strands, one each when $T = 3$, 10, 20, 50, and 100.

Figure 4 shows the rapid convergence of the distribution function to the large sample distribution function in the noninvertible case. This has a very long right-hand tail, while there is a large probability that the ML estimator is zero.

Figure 5 indicates that it takes quite a substantial sample size before the asymptotic normality result is useful. The probability that the ML estimator is zero is very persistent in this case, with a large sample size being required before this probability reduces substantially. When T is very small there is also a very considerable probability that the ML estimator is infinity.

6. THE FIRST-ORDER MOVING AVERAGE MODEL

The reduced form of the local level model is a constrained MA(1) imposing, on the representation (1.1), the constraint that $\theta \leq 0$. Although the development of Sections 2 through 5 gives the key to unlocking the general unconstrained MA(1) problem, there are a number of difficulties which have to be solved before this is actually achieved.

If y is written to denote $(y_1, \ldots, y_T)'$, then the log-likelihood can be expressed as

$$\begin{aligned}\log L(\sigma^2,\theta;y) &= \text{const} - \frac{T}{2}\log\sigma^2 - \frac{1}{2}\log|(1+\theta^2)I + \theta B| \\ &\quad - \frac{1}{2\sigma^2}\, y'((1+\theta^2)I + \theta B)^{-1}y \\ &= \text{const} - \frac{T}{2}\log\sigma^2 - \frac{1}{2}\sum_{t=1}^{T}\log\omega_t(\theta) \\ &\quad - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\sigma^{*2}u_t^2\,\frac{\omega_t(\theta^*)}{\omega_t(\theta)}, \end{aligned} \tag{6.1}$$

where θ^* and σ^{*2} denote the true values of θ and σ^2, respectively, $u_t \sim$ NID(0,1), and $\omega_t = (1+\theta^2) + \theta\lambda_t$. This allows the proof of Theorem 7.

THEOREM 7. *If the true value of σ^2 is strictly positive and the true value of $\theta \in (-1,1)$ and their ML estimators are written as $\hat{\theta}$ and $\hat{\sigma}^2$, then the following holds*:

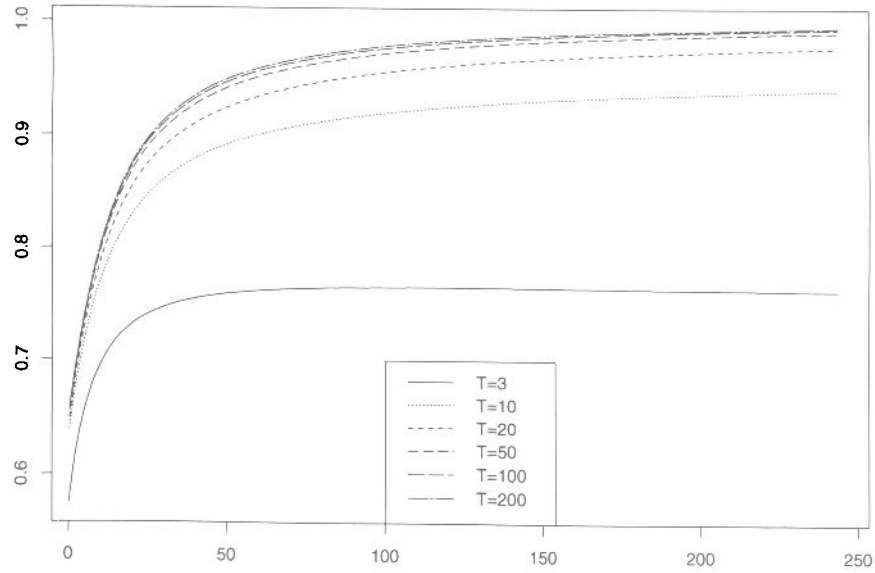

FIGURE 4. Approximate distribution function for $T^2\hat{q}$ when $q^* = 0$ for various values of T. X-axis: T^2q, Y-axis: distribution.

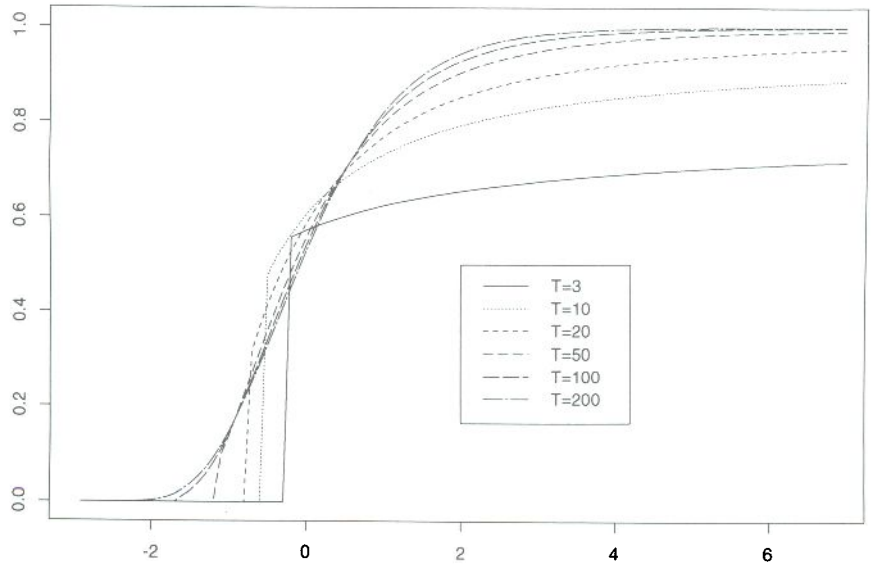

FIGURE 5. Approximate distribution function for standardized $\hat{q}$ when $q^* = 0.1$ for various values of T. X-axis: standardized q, Y-axis: distribution.

$$\sqrt{T}\begin{pmatrix}\hat{\sigma}^2-\sigma^2\\ \hat{\theta}-\theta\end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}2\sigma^4 & 0\\ 0 & 1-\theta^2\end{pmatrix}\right), \tag{6.2}$$

as the sample size goes to infinity.

Proof. Well known, see, for example, Harvey [17, p. 131]. ■

When the true value of θ is either 1 or -1, asymptotic normality fails as the model is then noninvertible. Some of the properties of the ML estimator under the noninvertibility conditions have been studied in Cryer and Ledolter [6], Sargan and Bhargava [32], Anderson and Takemura [2], and Tanaka and Satchell [35]. To make progress with the distribution theory, it is useful to concentrate the likelihood at

$$\hat{\sigma}^2 = \frac{y'((1+\theta^2)I+\theta B)^{-1}y}{T} \tag{6.3}$$

to give the profile likelihood function

$$\begin{aligned} M(\theta;y) &= \log L(\hat{\sigma}^2,\theta;y) \\ &= \text{const} - \frac{1}{2}\log|(1+\theta^2)I+\theta B| - \frac{T}{2}\log y'((1+\theta^2)I+\theta B)^{-1}y. \end{aligned} \tag{6.4}$$

The associated score is

$$\begin{aligned} s(\theta;y) &= \frac{dM(\theta;y)}{d\theta} \\ &= -\frac{1}{2}\operatorname{tr}(((1+\theta^2)I+\theta B))^{-1}(2\theta I+B)) \\ &\quad + \frac{T}{2}\frac{y'((1+\theta^2)I+\theta B)^{-1}(2\theta I+B)((1+\theta^2)I+\theta B)^{-1}y}{y'((1+\theta^2)I+\theta B)^{-1}y} \\ &= -\frac{1}{2}\sum_{t=1}^{T}\frac{2\theta+\lambda_t}{\omega_t(\theta)} + \frac{T}{2}\frac{\sum_{t=1}^{T}u_t^2\dfrac{(2\theta+\lambda_t)\omega_t(\theta^*)}{\omega_t^2(\theta)}}{\sum_{t=1}^{T}u_t^2\dfrac{\omega_t(\theta_t^*)}{\omega_t(\theta)}}. \end{aligned} \tag{6.5}$$

It is then a simple matter to show that $s(-1;y)$ and $s(1;y)$ are both exactly zero; a result which follows because the likelihood for θ is the same as the likelihood for θ^{-1}.

Equation (6.5) allows the proof of Theorem 8. Without loss of generality the true value of θ will be taken to be -1, for a mirror image property will deal with the case where it is 1.

THEOREM 8. *If the true value of* θ *is* -1 *and* θ *is written as* $-1 + c/T$, *then the following result holds*:

$$\frac{s(\theta;y)}{T} \xrightarrow{d} \frac{c}{2}\left[\sum_{t=1}^{\infty} u_t^2 \frac{t^2\pi^2}{(c^2 + t^2\pi^2)^2} - \sum_{t=1}^{\infty} \frac{1}{(c^2 + t^2\pi^2)}\right] = c\,r(c^2;u). \quad \textbf{(6.6)}$$

Proof. Tanaka and Satchell [35], Theorem 3. ■

Theorem 8 allows a straightforward proof of T-consistency of the ML estimator of θ. This is expressed formally in Theorem 9.

THEOREM 9. *The ML estimator of* θ *is T-consistent if the true value of* θ *is either* 1 *or* -1. *In the case of the value being* -1, *we mean that for any* $\epsilon > 0$ *there exists a* $c > 0$ *such that*

$$\lim_{T\to\infty} \operatorname{pr}\left(\frac{s(\theta;y)}{T}\bigg|_{\theta=-1+(c/T)} \geq 0\right) < \epsilon. \quad \textbf{(6.7)}$$

Proof. Theorem 4 of Tanaka and Satchell [35]. ■

Having recorded the speed of convergence of the estimator, attention now shifts to trying to derive the distribution of $T(\hat{\theta} + 1)$, where $\hat{\theta}$ denotes the ML estimator of θ. Although Tanaka and Satchell [35] proposed an approximation to this distribution, their results turned out to be too inaccurate for practical use. Instead, the approach suggested in Section 3 is adapted so that it is possible to use it here. The difficulty with the direct application of Theorem 6 is that the likelihood equation always has solutions at 1 and -1 in the unconstrained case.

Given a root at -1, it is possible to distinguish between it being a local maximum and a local minimum by inspecting the second derivative of the log-likelihood function. This has to be negative for there to be a maximum at -1, otherwise it will be a minimum. It is not difficult to show that

$$\begin{aligned}\frac{s'(-1;y)}{T} &= \frac{1}{T}\frac{ds(\theta;y)}{d\theta}\bigg|_{\theta=-1} \\ &= -\frac{1}{2T}\sum_{t=1}^{T}\frac{1}{(1-\lambda_t)} + \frac{1}{2}\frac{\sum_{t=1}^{T} u_t^2 \dfrac{\omega_t(\theta^*)}{(1-\lambda_t)^2}}{\sum_{t=1}^{T} u_t^2 \dfrac{\omega_t(\theta^*)}{(1-\lambda_t)}}. \end{aligned} \quad \textbf{(6.8)}$$

This allows the derivation of the results given in Theorem 10.

THEOREM 10. *If the true value of* θ *is written as* θ^*, *then the probability there exists a (local) maximum in the likelihood at* -1 *is*

$$\text{pr}(s'(-1;y) \leq 0) = \text{pr}\left[\sum_{t=1}^{T} u_t^2 \frac{\omega_t(\theta^*)}{1-\lambda_t}\left(\frac{1}{1-\lambda_t} - g\right) \leq 0\right],$$

$$\text{where} \quad g = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{1-\lambda_t}. \tag{6.9}$$

When $\theta^* = -1$, the probability that the ML estimator is exactly -1 is, in the limit,

$$\lim_{T\to\infty} \text{pr}(s'(-1;y) \leq 0) = \text{pr}\left[\sum_{t=1}^{\infty} u_t^2 \frac{1}{t^2\pi^2} \leq \frac{1}{6}\right] = 0.65744. \tag{6.10}$$

Proof: Sargan and Bhargava [32] and Tanaka and Satchell [35]. ■

Theorem 10 only captures one feature of the distribution of $T(\hat{\theta}+1)$. The key to establishing the rest of the distribution is Theorem 11, which parallels Theorem 5.

THEOREM 11. *Let λ denote a scalar parameter, $f(\lambda)$ some nonstochastic objective function, $s(\lambda)$ the derivative of f, $s'(\lambda)$ the second derivative of f, and $\hat{\lambda}$ the value of λ for which f is globally maximized over the range $[a,b]$. Assume:*

(i) *f is continuously twice differentiable for $\lambda \in [a,b]$,*
(ii) *$s(a) = s(b) = 0$,*
(iii) *$s'(a) \neq 0$ and $s'(b) \neq 0$,*
(iv) *there are a finite number of solutions to the equation $s(\bar{\lambda}) = 0$ and all these solutions are either maximums or minimums,*
(v) *there is exactly one point which is a local maximum in f.*

Then

$\hat{\lambda} = a$ if and only if $s'(a) < 0$,

$\hat{\lambda} = b$ if and only if $s'(b) > 0$,

$\hat{\lambda} \leq \lambda$ if and only if $s(\lambda) \leq 0, \quad \lambda \in (a,b)$. ■

This analytic result allows the proof of Theorem 12.

THEOREM 12. *Let λ denote a scalar parameter, $f(\lambda)$ some stochastic objective function, $s(\lambda)$ the derivative of f, $s'(\lambda)$ the second derivative of f, and $\hat{\lambda}$ the value of λ for which f is globally maximized over the range $[a,b]$. Assume:*

(i) *$s(\lambda)$ is continuous for $\lambda \in [a,b]$,*
(ii) *$s(a) = s(b) = 0$*

and U is written to denote the number of local maximums in f. Then

$$\text{pr}(\hat{\lambda} = a) - \text{pr}(s'(a) < 0) = \tau(a), \tag{6.11a}$$

$$\text{pr}(\hat{\lambda} \leq \lambda) - \text{pr}(s(\lambda) \leq 0) = \tau(\lambda), \quad \lambda \in (a,b), \tag{6.11b}$$

$$\text{pr}(\lambda \leq b) = 1, \tag{6.11c}$$

where

$$\tau(a) = \left(\text{pr}(\hat{\lambda} = a \mid U > 1) - \text{pr}(s'(a) < 0 \mid U > 1)\right) \text{pr}(U > 1), \tag{6.12a}$$

$$\tau(\lambda) = \left(\text{pr}(\hat{\lambda} \leq \lambda \mid U > 1) - \text{pr}(s(\lambda) < 0 \mid U > 1)\right) \text{pr}(U > 1), \quad \lambda \in (a,b). \tag{6.12b}$$

■

Theorem 12 suggests the use of the approximation

$$\text{pr}(\hat{\theta} = -1) \cong \text{pr}(s'(-1) < 0), \tag{6.13a}$$

$$\text{pr}(\hat{\theta} \leq \theta) \cong \text{pr}(s(\theta) \leq 0), \qquad \theta \in (-1,-1), \tag{6.13b}$$

$$\text{pr}(\hat{\theta} \leq 1) = 1, \tag{6.13c}$$

in this case. To assess $\tau(\theta;\theta^*)$, 5,000 replications of the log-likelihood were computed and searched for each combination of $\theta^* = -1$, -0.95, -0.9, -0.5, and 0 and $T = 3$, 10, 20, 50, 100, and 200. The estimate of $\text{pr}(U > 1)$ is given in Table 2 for these values. The reported probabilities are higher than the corresponding values in Table 1.

The corresponding plots for $\tau(\theta;\theta^*)$ are given in Figures 6 and 7 for $\theta^* = -1$ and 0, respectively. These consistently show τ having a substantial magnitude for $T = 3$ in both tails of the distribution. These results correspond to local, rather than global, maximums being estimated at both points of noninvertibility. When $\theta^* = -1$, $\tau(\theta)$ is small if θ is large and $T > 20$, while the left-hand tail diminishes less quickly. As $T \to \infty$, the precision of the approximation is exactly the same as that for the local level model. The reason for this is that

$$\lim_{T\to\infty} \text{pr}(T(\hat{\theta} + 1) \leq c) \cong \text{pr}\left(\sum_{t=1}^{\infty} u_t^2 \frac{t^2\pi^2}{(c^2 + t^2\pi^2)^2} - \sum_{t=1}^{\infty} \frac{1}{(c^2 + t^2\pi^2)} \leq 0\right) \tag{6.14}$$

$$\cong \lim_{T\to\infty} \text{pr}(T^2\hat{q} \leq c^2), \tag{6.15}$$

Table 2. Upper bound on the approximating error, $100 \times \text{pr}(U > 1)$

t	$\theta^* = -1$	$\theta^* = -0.95$	$\theta^* = -0.9$	$\theta^* = -0.5$	$\theta^* = 0$
3	5.16	5.18	5.14	4.50	4.08
10	3.42	3.42	3.30	3.80	5.02
20	1.64	1.60	1.58	2.12	2.06
50	1.00	1.16	0.98	0.88	0.20
100	0.94	0.88	1.04	0.12	0.02
200	0.98	1.06	0.76	0.00	0.00

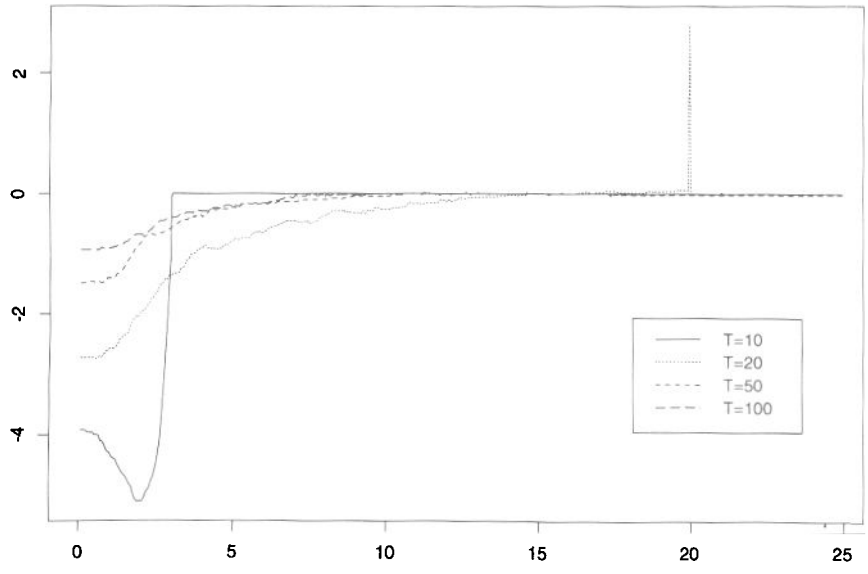

FIGURE 6. $\tau(\theta;\theta^* = -1)$ for various values of T. X-axis: $T(\theta + 1)$, Y-axis: 100τ.

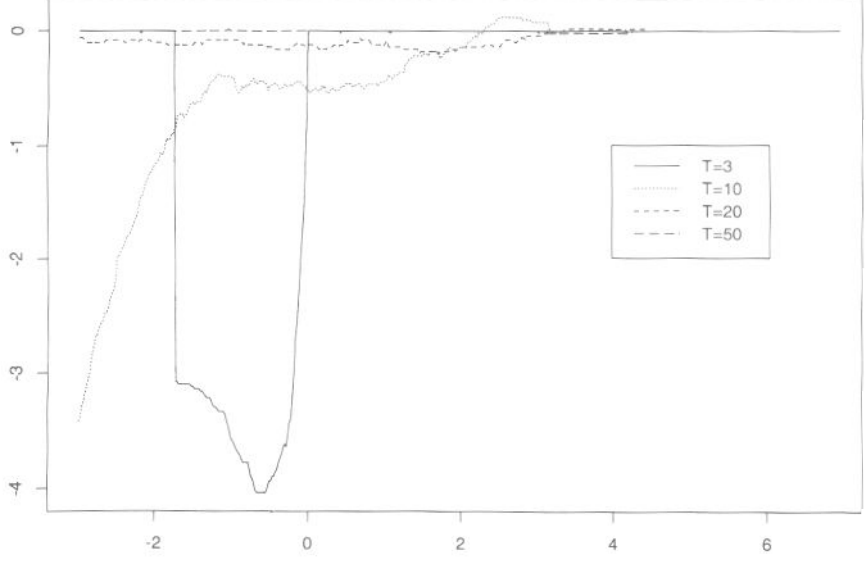

FIGURE 7. $\tau(\theta;\theta^* = 0)$ for various values of T. X-axis: standardized θ, Y-axis: 100τ.

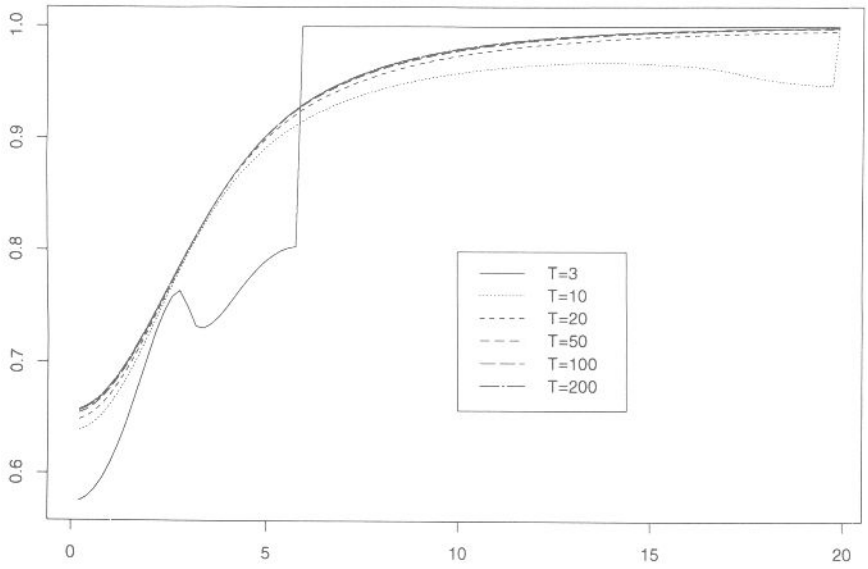

FIGURE 8. Approximate distribution function for $T(\hat{\theta} + 1)$ when $\theta^* = -1$ for various values of T. X-axis: $T(\theta + 1)$, Y-axis: distribution.

for the sign of the second derivative of the log-likelihood only depends on $r(c^2; u)$ when the sample size is large. The implication of this result is that the asymptotic distribution for the noninvertible MA(1) model can be read in Figure 4 by just square rooting the scale on the x-axis. The importance of this result is that the accuracy of this approximation is precisely the same as the accuracy in the local level model case. This was good, especially in the tails of the distribution function.

When $\theta^* > -1$, both probabilities of local estimates at the points of noninvertibility rapidly diminish, leaving quite an accurate approximation by the time T is 20. The speed of the convergence is dependent on θ^*, with θ^* being closest to zero converging quickest.

Generally, the graphs for τ suggest that the approximation given by (6.13) should be very good if $T > 20$. It is used to produce Figures 8 and 9 for the results corresponding to $\theta^* = -1.0$ and -0.75, respectively. The figures have six strands, one each for $T = 3$, 10, 20, 50, 100, and 200.

Figure 8 shows a very rapid convergence to the large sample distribution function in the noninvertible case. When $T = 3$, the magnitude of τ means that the approximated distribution function is not monotonically nondecreasing. Notice that it has a point mass of about 0.2 exactly on $\hat{\theta} = 1$, even though $\theta^* = -1$. However, by the time T has reached 10, τ is generally smaller than 0.02 in absolute value and so the function looks monotone.

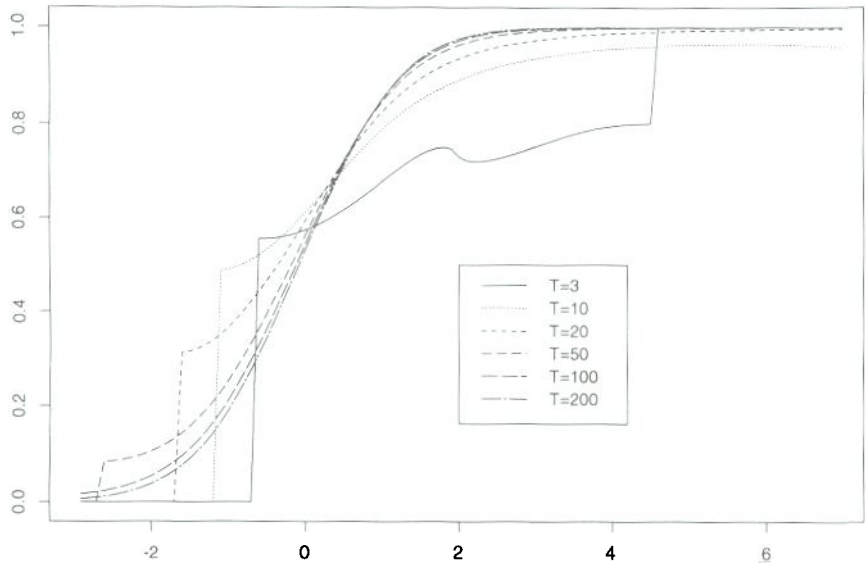

FIGURE 9. Approximate distribution function for standardized $\hat{\theta}$ when $q^* = -0.75$ for various values of T. X-axis: standardized θ, Y-axis: distribution.

Figure 9, which corresponds to θ^* being -0.75, has a smaller kink. The large sample asymptotics are very poor in this case, even when $T = 100$ due to the persistence in the estimation of noninvertibility. A further implication in this figure is that the asymptotic distribution severely underestimates the extreme right-hand tail of the distribution of the estimator.

7. CONCLUSION

This paper has presented approximations for the distribution function of the ML estimators of a local level model, which is a constrained MA(1) process and an unconstrained MA(1). The key results are given in equations (4.1) and (6.13).

The approximations are quite accurate if T is below 20, and very accurate for T above 50. When the moving average is noninvertible, the approximation does not hold exactly when $T \to \infty$. The error that results from using it is, however, very small, especially in the important right-hand tail of the distribution.

REFERENCES

1. Anderson, T.W. *The Statistical Analysis of Time Series.* New York: Wiley, 1971.
2. Anderson, T.W. & A. Takemura. Why do noninvertible estimated moving averages occur? *Journal of Time Series Analysis* 7 (1986): 235–254.

3. Ansley, C.F. & P. Newbold. Finite sample properties of estimators for autoregressive moving average models. *Journal of Econometrics* 13 (1980): 159–183.
4. Barndorff-Nielson, O.E. & D.R. Cox. *Asymptotic Techniques for Use in Statistics.* London: Chapman and Hall, 1989.
5. Cooper, D.M. & R. Thompson. A note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika* 64 (1977): 625–628.
6. Cryer, J.D. & J. Ledolter. Small-sample properties of the maximum likelihood estimator in the first-order moving average model. *Biometrika* 68 (1981): 691–694.
7. Daniels, H.E. Saddlepoint approximations for estimating equations. *Biometrika* 70 (1983): 89–96.
8. Davidson, J. Small sample properties of estimators of the moving average process. In E.G. Charatsis (ed.), *Proceedings of the Econometric Society Meeting 1979: Selected Econometric Papers in Memory of Stephan Valananis.* Amsterdam: North-Holland, 1979.
9. Davidson, J. Problems with the estimation of moving average processes. *Journal of Econometrics* 16 (1981): 295–310.
10. Davies, R.B. Numerical inversion of a characteristic function. *Biometrika* 60 (1973): 415–417.
11. Davies, R.B. AS 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics* 29 (1980): 323–333.
12. Dent, W. & A-S. Min. A Monte Carlo study of autoregressive integrated moving average processes. *Journal of Econometrics* 7 (1978): 23–55.
13. Farebrother, R.W. AS 153: Pan's procedure for the tail probabilities of the Durbin-Watson statistic. *Applied Statistics* 29 (1980): 224–227.
14. Field, C.A. & F.R. Hampel. Small-sample asymptotic distributions of *M*-estimators of location. *Biometrika* 69 (1982): 29–46.
15. Grenander, U. & G. Szego. *Toeplitz Forms and Their Applications.* Berkeley, CA: University of California Press, 1958.
16. Hannan, E.J. *Multiple Time Series.* New York: Wiley, 1970.
17. Harvey, A.C. *Time Series Models.* Oxford: Philip Allen Publishers Limited, 1981.
18. Harvey, A.C., B. Henry, S. Peters, & S. Wren-Lewis. Stochastic trends in dynamic regression models: An application to the employment-output equation. *Economic Journal* 96 (1986): 975–985.
19. Harvey, A.C. & J.H. Stock. Continuous time autoregressive models with common stochastic trends. *Journal of Economic Dynamics and Control* 12 (1988): 365–384.
20. Harvey, A.C. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.
21. Harville, D.A. Bayesian inference for variance components using only error contrasts. *Biometrika* 61 (1974): 383–385.
22. Huber, P.J. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35 (1964): 73–101.
23. Huzurbazar, V.S. The likelihood equation, consistency and the maxima of the likelihood function. *Annals of Eugenics* 14 (1948): 185–200.
24. Kalbfleisch, J.D. & D.A. Sprott. Applications of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society, Series B* 32 (1970): 175–194.
25. Kang, K.M. *A Comparison of Estimators of Moving Average Processes*, unpublished, from the Australian Bureau of the Census.
26. McCullagh, P. & J.A. Nelder. *Generalized Linear Models, 2nd ed.* London: Chapman and Hall, 1989.
27. Muth, J.F. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55 (1960): 299–305.
28. Pesaran, M.H. A note on the maximum likelihood estimation of regression models with first order moving average errors with roots on the unit circle. *Australian Journal of Statistics* 25 (1983): 442–448.

29. Pötscher, B.M. Noninvertibility and pseudo maximum likelihood estimation of misspecified ARMA models. *Econometric Theory* 7 (1991): 435–449.
30. Robinson, D.L. Estimation and use of variance components. *The Statistician* 36 (1987): 3–14.
31. Robinson, P.M. Robust nonparametric autoregression. In J. Frenke, W. Härdie, & D. Martin (eds.), *Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics*, Vol. 26, pp. 247–255. New York: Springer-Verlag, 1984.
32. Sargan, J.E. & A. Bhargava. Maximum likelihood estimation of regression models with first order moving average errors when the root lies on the unit circle. *Econometrica* 51 (1983): 799–820.
33. Shephard, N. Maximum likelihood estimation of regression models with stochastic trend components. *Journal of the American Statistical Association* 88 (1993): forthcoming.
34. Shephard, N. & A.C. Harvey. On the probability of estimating a deterministic component in the local level model. *Journal of Time Series Analysis* 11 (1990): 339–347.
35. Tanaka, K. & S.E. Satchell. Asymptotic properties of the maximum likelihood and nonlienar least squares estimator for noninvertible moving average models. *Econometric Theory* 5 (1989): 333–353.
36. West, M. & P.J. Harrison. *Bayesian Forecasting and Dynamic Models.* New York: Springer-Verlag, 1989.
37. White, J.S. The limiting distribution of the serial correlation coefficient in the explosive case. *Annals of Mathematical Statistics* 29 (1958): 1188–1197.

APPENDIX

Proof of Theorem 1. The asymptotic variance will be given by the inverse of the limiting average information per observation matrix. However,

$$\frac{\partial^2 \log L}{\partial(\sigma^2)^2} = \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{T}\sigma^{*2}u_t^2\frac{\delta_t(q^*)}{\delta_t(q)}, \tag{A.1}$$

$$\frac{\partial^2 \log L}{\partial\sigma^2\,\partial q} = -\frac{1}{2\sigma^4}\sum_{t=1}^{T}\sigma^{*2}u_t^2\frac{\delta_t(q^*)}{\delta_t(q)^2}, \tag{A.2}$$

$$\frac{\partial^2 \log L}{\partial q^2} = -\frac{1}{\sigma^2}\sum_{t=1}^{T}\sigma^{*2}u_t^2\frac{\delta_t(q^*)}{\delta_t^3(q)} + \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\delta_t^2(q)}. \tag{A.3}$$

Evaluating q and σ^2 at their true values and taking expectations means that we need to look at the quantities

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\delta_t^r(q)} = \frac{1}{\pi}\int_0^{\pi}\frac{1}{(4\sin^2\lambda/2 + q)^r}\,d\lambda + O(T^{-1}) \tag{A.4}$$

(see Grenander and Szego [15, p. 221, equation 7] or Hannan (16, p. 353].

$$= \frac{(-1)^{r-1}}{(r-1)!}\frac{d^{r-1}(q(q+4))^{-1/2}}{dq^{r-1}} + O(T^{-1}). \tag{A.5}$$

This implies the limiting average information per observation is given by

$$\frac{1}{2}\begin{bmatrix} \sigma^{-4} & \dfrac{1}{\sigma^2\sqrt{q(q+4)}} \\ \dfrac{1}{\sigma^2\sqrt{q(q+4)}} & \dfrac{(q+2)}{(q(q+4))^{3/2}} \end{bmatrix}. \tag{A.6}$$

Hence, by using a matrix inverse, the stated result is obtained. ■

Proof of Theorem 2. Think about

$$X(q) = \frac{1}{T}\sum_{t=1}^{T} u_t^2 \frac{\delta_t(0)}{\delta_t(q)}. \tag{A.7}$$

Recall

$$\delta_t(q) = 2 + q - 2\cos\frac{t\pi}{(T+1)} \tag{A.8}$$

$$= q + 4\sin^2\frac{t\pi}{2(T+1)} \tag{A.9}$$

$$= q + \frac{t^2\pi^2}{T^2} + O(T^{-4}). \tag{A.10}$$

So

$$X\left(\frac{c}{T^2}\right) = \frac{1}{T}\sum_{t=1}^{T} u_t^2 \frac{T^2\delta_t(0)}{T^2\delta_t(c/T^2)} = \frac{1}{T}\sum_{t=1}^{T} u_t^2 \frac{t^2\pi^2}{t^2\pi^2 + c} + O_p(T^{-1}) \xrightarrow{p} 1. \tag{A.11}$$

Likewise

$$Y\left(\frac{c}{T^2}\right) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\delta_t(c/T)} \to \sum_{t=1}^{\infty} \frac{1}{t^2\pi^2 + c} \tag{A.12}$$

and

$$Z\left(\frac{c}{T^2}\right) = \frac{1}{T}\sum_{t=1}^{T} u_t^2 \frac{\delta_t(0)}{\delta_t^2(c/T)} = \frac{1}{T}\sum_{t=1}^{T} u_t^2 \frac{t^2\pi^2}{(t^2\pi^2 + c)^2} + O_p(T^{-1})$$

$$\xrightarrow{d} \sum_{t=1}^{\infty} u_t^2 \frac{t^2\pi^2}{(t^2\pi^2 + c)^2}. \tag{A.13}$$

This proves the desired result. ■

Proof of Theorem 3. This follows the style of proof of Proposition 1 of Tanaka and Satchell [35]. Write

$$r(c) = \sum_{t=1}^{\infty} u_t^2 \frac{t^2\pi^2}{(t^2\pi^2 + c)^2} - \sum_{t=1}^{\infty} \frac{1}{t^2\pi^2 + c}, \tag{A.14}$$

which is a scaled version of limit of the score evaluated at c/T^2. The probability the score is positive can be made arbitrarily close to zero by selecting c. Straightforwardly

$$\mathrm{Er}(c) = -c \sum_{t=1}^{\infty} \frac{1}{(t^2\pi^2 + c)^2} \tag{A.15}$$

$$\mathrm{Var}(c) = \sum_{t=1}^{\infty} \frac{t^4\pi^4}{(t^2\pi^2 + c)^4}, \tag{A.16}$$

which implies by Chebyshev's inequality

$$\begin{aligned}\mathrm{pr}(r(c) \geq 0) &< \frac{\mathrm{Var}(c)}{(\mathrm{Er}(c))^2} \\ &= \frac{c^{-2} \sum_{t=1}^{\infty} \frac{t^4\pi^4}{(t^2\pi^2 + c)^4}}{\left(\sum_{t=1}^{\infty} \frac{1}{(t^2\pi^2 + c)^2} \right)^2} \\ &= c^{-2} g(c). \end{aligned} \tag{A.17}$$

$g(c)$ is such that for all c there exists a b such that $g(c) \leq b$, so setting $c = \sqrt{b/\epsilon}$ gives the result immediately. ■

Proof of Theorem 4. A formal Taylor series expansion of

$$\begin{aligned} r(\hat{c};u) = 0 = r(0;u) &+ \frac{dr(0;u)}{dc}(\hat{c} - 0) + \frac{d^2 r(0;u)}{dc^2} \frac{(\hat{c} - 0)^2}{2} + \dots \\ &= \sum_{t=1}^{\infty} \frac{u_t^2 - 1}{t^2\pi^2} - \hat{c} \sum_{t=1}^{\infty} \frac{2u_t^2 - 1}{(t^2\pi^2)^2} + \hat{c}^2 \sum_{t=1}^{\infty} \frac{3.2u_t^2 - 2}{2(t^2\pi^2)^3} \\ &\quad - \hat{c}^3 \sum_{t=1}^{\infty} \frac{4!u_t^2 - 3!}{3!(t^2\pi^2)^4} + \hat{c}^4 \sum_{t=1}^{\infty} \frac{5!u_t^2 - 4!}{4!(t^2\pi^2)^5} - \dots . \end{aligned} \tag{A.18}$$

Canceling the factorials gives the result in the theorem. The formal expansion will be valid provided $\hat{c} \in (-\pi^2, \pi^2)$, for then the expectation of the sum is bounded. Outside that range this is not the case. ■

Proof of Theorem 5. By (ii), $s(a)$ has to be positive or negative. First, let us deal with the case where it is negative. There is a local maximum at $\lambda = a$ and so there cannot exist another local maximum by (v). f must continually fall unless there is a local minimum by (iv). There cannot exist a point which is a local minimum, as this would create a local maximum. The only exception to this is if $s(b) = 0$, which is itself excluded by (iii). Hence, if $s(a) < 0$, then $s(\lambda) < 0 \forall \lambda \in [a, b]$, and $\hat{\lambda} = a$.

If $s(a)$ is positive, then it will continue being positive until it falls through zero. This point is the unique local maximum. f must then fall continually using the same arguments as before. ■

Proof of Theorem 6

$$\begin{aligned}\mathrm{pr}(\hat{\lambda} \leq \lambda) &= \mathrm{pr}(\hat{\lambda} \leq \lambda \,|\, U = 1)\, \mathrm{pr}(U = 1) + \mathrm{pr}(\hat{\lambda} \leq \lambda \,|\, U > 1)\, \mathrm{pr}(U > 1) \\ &= \mathrm{pr}(s(\lambda) \leq 0 \,|\, U = 1)\, \mathrm{pr}(U = 1) + \mathrm{pr}(\hat{\lambda} \leq \lambda \,|\, U > 1)\, \mathrm{pr}(U > 1) \end{aligned} \tag{A.19}$$

by Theorem 5. As a result,

$$\text{pr}(\hat{\lambda} \leq \lambda) - \text{pr}(s(\lambda) \leq 0)$$
$$= [\text{pr}(\hat{\lambda} \leq \lambda \mid U > 1) - \text{pr}(s(\lambda) \leq 0 \mid U > 1)]\text{pr}(U > 1). \quad \blacksquare$$

Proof of Theorem 11. Follows the same lines as the proof of Theorem 5. The complication is that we start with $s(a) = 0$, so now the second derivative of f has to be inspected. By (iii), $s'(a)$ has to be positive or negative. Let us deal with the case where it is negative. There is a local maximum at $\lambda = a$, and so there does not exist another local maximum by (v). f must continually fall unless there is a local minimum by (iv). There does not exist a local minimum, as this would create a local maximum. The only exception to this is that $s(b) = 0$ and so $\lambda = b$ must be a minimum by (v). Hence, if $s'(a) < 0$, then $\hat{\lambda} = a$.

The rest of the result follows using the same argument.

Proof of Theorem 12. Follows exactly the lines of Theorem 6 but now using Theorem 11's analytic result rather than Theorem 5's. $\blacksquare$