Econometric Theory, **15**, 1999, 24–49. Printed in the United States of America.

ASYMPTOTICS OF ML ESTIMATOR FOR REGRESSION MODELS WITH A STOCHASTIC TREND COMPONENT

BIING-SHEN KUO
National Chengchi University

This paper investigates the asymptotic properties of the maximum marginal likelihood estimator for a regression model with a stochastic trend component when the signal-to-noise ratio is near zero. In particular, the local level model in Harvey (1989, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press) and its variants where a time trend or an intercept is included are considered. A local-to-zero parameterization is adopted. Two sets of asymptotic properties are presented for the local maximizer: consistency and the limiting distribution. The estimator is found to be super-consistent. The limit distribution is derived and found to possess a long tail and a mass point at zero. It yields a good approximation for samples of moderate size. Simulation also documents that the empirical distribution converges less rapidly to the limit distribution as number of regression parameters increases. The results could be viewed as a transition step toward establishing new likelihood ratio–type or Wald-type tests for the stationarity null.

1. INTRODUCTION

Structural time series modeling in econometrics is popular and useful. In a structural model of an economic system, each component is intended to represent a specific observable feature of the data such as trends or seasonals. The simplest form of the structural model is the local level model of Harvey (1989) where the time series under study is written as the sum of a stochastic trend and an error, $y_t = \gamma_t + u_t$, $\gamma_t = \gamma_{t-1} + v_t$. Here $u_t \sim \text{nid}(0,\sigma^2)$, $v_t \sim \text{nid}(0, q\sigma^2)$ with $q \geq 0$, called the signal-to-noise ratio, and the random walk component is to permit the structure to evolve over time. Use of the model can be traced back to Muth (1960), who showed that it provides a rationale for forecasting schemes that place more weight on the most recent observations. Cooley and Prescott (1976) studied the estimation of a time-varying coefficient regression on the basis of the model. Its prac-

I am deeply indebted to Serena Ng, two referees, and a co-editor for helpful comments and suggestions that led to a great improvement of the paper, to Richard Davis for stimulating correspondence, and especially to Bruce Hansen for offering excellent advice while a previous version of the article was written under his supervision as a part of my doctoral thesis at the University of Rochester. The financial support from the Chiang Ching Kuo Foundation through the dissertation fellowship is also gratefully acknowledged. Address correspondence to: Biing-Shen Kuo, Graduate Institute of International Trade, National Chengchi University, Taipei 116, Taiwan; e-mail: bsku@nccu.edu.tw.

 © 1999 Cambridge University Press 0266-4666/99 $9.50

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

tical appeal has also led to elaborate applications, such as Harvey, Henry, Peters, and Wren-Lewis (1986) on the employment-output equation and Harvey and Stock (1988) on the income–consumption relationship. Recently this sort of model has been applied to unit root testing. As noted in Schmidt and Phillips (1992), the framework gives rise to a natural interpretable parameterization of the permanent and transitory components. Having observed that the hypothesis that the variance of the random walk component equals zero corresponds to the hypothesis of stationarity, Kwiatkowski, Phillips, Schmidt, and Shin (1992) proposed LM-type tests for the null of stationarity, in contrast to the Dickey–Fuller-type test having a unit root null.

Although the model is useful in practical analysis, the asymptotic theory for it has not been thoroughly studied. In particular, the limit distribution of the maximum likelihood estimator for such a simple model when $q = 0$ has remained unknown. It appears intractable to develop the Wald test and the likelihood ratio (LR) test for the stationarity null because both tests entail estimation of the model under the null and the alternative. Yet the development for these tests is important. In the context where the parameter of concern lies on the boundary under the null, the Wald, LR, and LM tests are not asymptotically equivalent. Hence, it is reasonable to expect that the Wald and LR tests possess better power against some alternatives than the LM test. More importantly, whether q is zero suggests the integration order of the series, giving different empirical implications for assessing forecasting performance and certain economic theories. Because the asymptotic normal distribution in the boundary case yields a poor approximation to the actual one, inference based on it may actually be inadequate. The interest in obtaining a comprehensive limit theory for the model is then well deserved.

The difficulty in doing so, however, lies in the nonstandard nature of the problem. When the true q equals zero and thus is at the boundary of the parameter space, the conventional Taylor series approach can not be immediately employed to obtain the limit distribution. Virtually, as will be shown later, an approximately constant information matrix can not be found around the true parameter, which violates one of the regularity conditions required in the standard limit distribution argument (see, e.g., Amemiya, 1985). The problem is further compounded by the observation that q is frequently estimated to be zero (see Shephard and Harvey, 1990). It indicates that there is a fairly complicated distribution function in the limit.

This paper investigates the asymptotic properties for the local level model and its variants with q being near or at zero. The results presented build on the recent asymptotic theory by Davis and Dunsmuir (1996) (D&D hereafter) for MA(1) process with a root on or near the unit circle. Following a differencing procedure, it is easy to see that the local level model is equivalent to an IMA(1,1) model with the moving average parameter θ ($\in[-1,1]$) related to q by $q = (1 - \theta)^2/\theta$. The correspondence reflects that derivation of the asymptotics for the MA(1) model is as irregular as that of the local level model. The usual asymptotic theory for ML estimator would break down again when θ is on the unit circle.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

The noninvertible MA(1) case with $\theta = 1$ has indeed received considerable attention. Sargan and Bhargava (1983), Anderson and Takemura (1986), and Tanaka and Satchell (1989) are a partial list for previous analytical work on the case. It was not until D&D that the limit theory was completed. On the other hand, as $q = 0$ if and only if $\theta = 1$, we see the equivalence between testing for the stationarity null and that for a moving average unit root. Saikkonen and Luukkonen (1993) embodied this notion in their tests, motivated by a local optimality argument.

The large-sample theory derived in this paper can be viewed as extensions of the D&D theory. First, our attention is given not only to the simple local level model but also to the model with an intercept and a deterministic time trend. These are common characterizations for macroeconomic variables and hence are worth special consideration. Although applicable to a broader class of models, D&D's theory for the MA(1) model can not readily lend itself to deriving the asymptotics for these structural models. Our results, in this way, may constitute a contribution to filling up the gap. A further extension is made by considering another class of estimators. When additional regression parameters are involved, as in the models studied here, one would typically use an ordinary ML estimator with a concentrated likelihood function to eliminate the nuisance parameters. Its use in this context, unfortunately, has an undesirable property that q is estimated to be zero with a high probability even when the true q is nonzero (see Shephard and Harvey, 1990). To cope with the property, we adopt in the paper the maximum marginal likelihood (MML) estimator. The MML estimator is commonly suggested in the previous literature (e.g., Shephard, 1993a) as it has the merit of having a zero expectation of its score. In other words, the MML estimator can give unbiased estimates for q. Without this merit, the ML estimator is obviously less preferable. Nevertheless, the results with the ML estimator is also provided here to permit a comparison.

Our findings are consistent with those in Shephard and Harvey (1990) and Shephard (1993b). The asymptotic properties of θ and q are quite distinct because the former is found to be T-consistent on the unit circle, whereas the latter is T^2-consistent on the zero boundary. Of more practical relevance is that for samples of moderate size, the limit distribution proves to well approximate the empirical counterpart, featured by a very long right-hand tail and a jump at zero with a high probability. We conclude that our results are of potential use in constructing new Wald or LR tests for the stationarity null.

The remainder of the paper is organized as follows. Section 2 lists structural models of concern in the paper and simplifies the likelihood using the latent structure under a local-to-zero parameterization. Section 3 establishes two sets of asymptotic properties for the local maximizer estimator, consistency and the limit distribution. The accuracy of the limit approximation is investigated in Section 4 for a variety of sample sizes and parameter values. Section 5 concludes. The Appendix contains the proofs.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

2. A REGRESSION MODEL WITH STOCHASTIC TREND COMPONENT AND ITS LIKELIHOOD

2.1. The Model

The analysis begins with a component model where there are an intercept, a deterministic trend, a stochastic trend, and a stationary noise:

$$y_t = \alpha + \beta t + \gamma_t + u_t \tag{1}$$

and

$$\gamma_t = \gamma_{t-1} + v_t, \qquad \gamma_0 = \gamma, \qquad t = 1, \ldots, T, \tag{2}$$

where $u_t \sim \text{nid}(0, \sigma^2)$ and $v_t \sim \text{nid}(0, q\sigma^2)$ which is independent of u_s for all of t and s. In the model σ^2 is a scale parameter that does not affect the estimation of α and β, whereas q is the signal-to-noise ratio that determines the properties of the component model. The focus of the paper is on the asymptotic properties of maximum likelihood estimation for q when its true value, denoted q^*, is on or near the boundary of zero. Thus together with σ^2, α and β are the nuisance parameters to be estimated. The random walk process, γ_t, is unobservable and starts with γ_0 taken to be a fixed known or unknown constant. Because it can not be separately identified from α, γ_0 is assumed to be zero.

We also consider the model without a time trend given by

$$y_t = \alpha + \gamma_t + u_t. \tag{3}$$

It is easily seen that the local level model, $y_t = \gamma_t + u_t$, can be embraced as a subcase of (3) where α is now replaced by the startup of the stochastic trend process, γ.

In combination with (2), the regression models (1) and (3) can be expressed as

$$y_t = x_t' \delta + \epsilon_t,$$

where, respectively, $x_t = (1, t)'$ or a unity, $\delta = (\alpha, \beta)'$ or α, and $\epsilon_t = u_t + \sum_1^t v_i$; or in compact matrix form,

$$y = X\delta + \epsilon,$$

where y is a $T \times 1$ vector containing the y_t, X is a matrix with tth row vector equal to x_t', and ϵ is a $T \times 1$ vector containing the ϵ_t. The y is thus normally distributed as

$$y \sim N(X\delta, \sigma^2 \Omega(q)), \qquad \Omega(q) = I_T + qQ, \tag{4}$$

where $Q = LL'$ with $L = ((I(i \geq j))$ and $I(\cdot)$ is the indicator function.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

2.2. Marginal Likelihood

Typically inference about q is based on the maximum likelihood estimation

$$\max_{\delta,q,\sigma^2} -\frac{T}{2}\log\sigma^2 - \frac{1}{2}\log|\Omega(q)| - \frac{(y-X\delta)'\Omega(q)^{-1}(y-X\delta)}{2\sigma^2}.$$

To reduce the dimensionality, we shall derive the concentrated likelihood by conditioning δ and σ^2 out of the function. Maximizing the likelihood partially with respect to δ and σ^2 conditional on q yields the estimators

$$\hat{\delta}(q;y) = (X'\Omega(q)^{-1}X)^{-1}X'\Omega(q)^{-1}y, \qquad \hat{\sigma}^2(q;y) = \frac{(y-X\hat{\delta})'\Omega(q)^{-1}(y-X\hat{\delta})}{T}.$$

These are plugged into the likelihood to determine the concentrated likelihood:

$$L_k^p(q;P_\Omega y) = -\frac{1}{2}\log|\Omega(q)| - \frac{T}{2}\log y'P_\Omega(q)\Omega(q)^{-1}P_\Omega(q)y,$$

where the projection matrix $P_\Omega(q) = I_T - X(X'\Omega(q)^{-1}X)^{-1}X'\Omega(q)^{-1}$, k is the dimension of δ taking value 1 or 2, and the constant term is ignored. The concentrated likelihood thus no longer depends on δ and σ^2.

The term $L_k^p(q;P_\Omega y)$ is known as the profile likelihood. Using it, though offering numerical convenience, also has disadvantages. Most notably, it is not a usual log likelihood function whose derivative does not have a zero mean. This simply results from the fact that $\delta(\cdot)$ and $\sigma^2(\cdot)$, depending on the entire vector y, enter $L_k^p(\cdot)$ as arguments.[1]

When using the profile likelihood in the current context, the feature with a nonzero mean score leads to a undesirable property. As demonstrated in Shephard and Harvey (1990) and Shephard (1993a), the estimates for q are found to be zero with a considerable nonzero probability even when q^* is nonzero. They suggested using the marginal likelihood to mitigate the problem. The marginal likelihood for q is based on the probability density of the invariant statistic $P_\Omega(q)y$, instead of y. For our models, after concentrating out σ^2, the marginal likelihood is (see McCullagh and Nelder, 1989)

$$L_k^m(q;P_\Omega y) = -\frac{1}{2}\log|\Omega(q)| - \frac{1}{2}\log|X'\Omega(q)^{-1}X| - \frac{T-k}{2}\log y'P_\Omega'(q)\Omega(q)^{-1}P_\Omega(q)y,$$

which differs from the profile likelihood by the term $\frac{1}{2}\log|X'\Omega(q)^{-1}X|$.

Although both likelihoods are driven by the same random variable, $P_\Omega(q)y$, the marginal likelihood is clearly preferable to the profile likelihood. This is because the former can yield a zero mean of its score and hence can give unbiased estimates for q. This is a property that is essential in estimation. Indeed, it is the bias

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

that makes acute the problem of estimating q around the boundary. In what follows, we confine our attention to the marginal likelihood when deriving the asymptotics.

The procedures to estimate q require repeated evaluations of L_k^m, its score and derivative of the score. For practical purposes and asymptotic analysis, it is worthwhile to simplify the expression of L_k^m, in particular the quadratic term $y'P_\Omega'(q)\Omega(q)^{-1}P_\Omega(q)y$. We express the likelihood in terms of latent roots. This is done by using the following relation derived in Patterson and Thompson (1971),

$$y'P_\Omega'(q)\Omega(q)^{-1}P_\Omega(q)y = y'W\,\mathrm{diag}((1+q\lambda_{s,T-k})^{-1})W'y,$$

where W is a $T\times(T-k)$ matrix whose columns are orthogonal vectors to $P = P_\Omega(0) = I_T - X(X'X)^{-1}X$ with the properties: (a) $W'W = I_{T-k}$, (b) $WW' = P$, and (c) $W'\Omega(q)W = \mathrm{diag}(1+q\lambda_{s,T-k})$, a diagonal matrix with elements $1+q\lambda_{s,T-k}(s=1,\ldots,T-k)$. From property (c), the $\lambda_{s,T-k}$ are the nonzero latent roots of $W'QW$, a result of $(W'\Omega(q)W)W' = \mathrm{diag}(1+q\lambda_{s,T-k})W'$. Because $W'y \sim N(0,\sigma^2(I_{T-k}+q^*W'QW))$, the quadratic term is then

$$y'P_\Omega'(q)\Omega(q)^{-1}P_\Omega(q)y = \sum_{s=1}^{T-k}\frac{1+q^*\lambda_{s,T-k}}{1+q\lambda_{s,T-k}}\sigma^2 w_s^2, \qquad w_s \sim \mathrm{nid}(0,1). \tag{5}$$

Note that the latent roots of $W'QW$, $\lambda_{s,T-k}$, are indexed by the regressor number. When $k=2$ in (1), the $(T-2)$ nonzero latent roots are given in Nyblom (1986) as

$$\lambda_{2s-1,T-2} = \frac{1}{2(1-\cos(2s\pi/T))}, \qquad \lambda_{2s,T-2} = \frac{1}{2(1-\cos\phi_{s,T-2})},$$
$$s = 1,\ldots,\frac{T-2+\mathrm{Re}(T,2)}{2},$$

where $\mathrm{Re}(T,2)$ is the remainder of $T/2$ and $\phi_{s,T-2}\in(2s\pi/T,(2s+1)\pi/T)$ such that $T\tan(\phi/2) = \tan(T\phi/2)$. In the case that $k=1$ in (3), Nyblom and Mäkeläinen (1983) found that $(T-1)$ latent roots to be

$$\lambda_{s,T-1} = \frac{1}{2(1-\cos(s\pi/T))}, \qquad s = 1,\ldots,T-1.$$

Also for future use, the latent roots of Q take the form

$$\lambda_{s,T} = \frac{1}{2(1-\cos[(2s-1)\pi/(2T+1)])}, \qquad s = 1,\ldots,T.$$

The sum of two remaining deterministic terms in the likelihood, $\log|\Omega(q)| + \log|X'\Omega(q)^{-1}X|$, can also be written in terms of the latent roots of $W'QW$. This is based on an equality proven in the Appendix that

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

$$\sum_{s=1}^{T-k} \log(1 + q\lambda_{s,T-k}) = \log|\Omega(q)| + \log|X'\Omega(q)^{-1}X| + \text{Constant}. \quad \textbf{(6)}$$

Adding (5) and (6) together, apart from a constant term, the marginal likelihood can now be rewritten as

$$L_k^m(q) = -\frac{1}{2}\sum_{s=1}^{T-k} \log(1 + q\lambda_{s,T-k}) - \frac{T-k}{2}\log \sum_{s=1}^{T-k} \frac{1 + q^*\lambda_{s,T-k}}{1 + q\lambda_{s,T-k}}\,\sigma^2 w_s^2.$$

Note that when $k = 1$, the likelihood is equivalent to that of the differenced local level model studied by Shephard (1993b).

2.3. The Score and Its Derivative

Because inference on q when the true value is 0 or near 0 is of primary interest, a local-to-zero parameterization is adopted:

$$q = \frac{d}{T^2},$$

where T is the sample size and $d \geq 0$. The normalization factor, T^2, is taken to match the asymptotics. Making the parameter dependent on the sample size is common in the statistical literature (e.g., the Pitman sequence) and in the recent unit root literature. The motivation is to deliver a useful approximation for finite samples when the true parameter is on or near the boundary. The formulation could have the interpretation that the estimator for q and hence d would not be consistent unless the amount of "local" information on which it depends accumulates in a suitable way. That is, an increase in the length of the series T does not imply an improvement in estimation of d.

The marginal likelihood with true signal-to-noise ratio $q^* = d^*/T^2$ now expressed as a function of d is

$$l_{k,T}(d) = L_k^m(q).$$

In the next section, we are mainly concerned with the asymptotics of the score and its derivative. Their sample counterparts are

$$\begin{aligned} l'_{k,T}(d) &= \left(\frac{\partial L_k^m(q)}{\partial q}\right)\left(\frac{\partial q}{\partial d}\right) \\ &= \frac{1}{2T^2}\left\{(T-k)g_{k,T} - \sum_{s=1}^{T-k} \frac{1}{d/T^2 + \lambda_{s,T-k}^{-1}}\right\}, \end{aligned}$$

where

$$g_{k,T} = \left(\sum_{s=1}^{T-k} \frac{d^*/T^2 + \lambda_{s,T-k}^{-1}}{d/T^2 + \lambda_{s,T-k}^{-1}}\, w_s^2\right)^{-1}\left(\sum_{s=1}^{T-k} \frac{d^*/T^2 + \lambda_{s,T-k}^{-1}}{(d/T^2 + \lambda_{s,T-k}^{-1})^2}\, w_s^2\right)$$

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

and

$$l''_{k,T}(d) = \left(\frac{\partial^2 L_k^m(q)}{\partial q^2}\right)\left(\frac{\partial q}{\partial d}\right)^2$$
$$= -\frac{1}{2T^4}\left\{(T-k)g'_{k,T} - \sum_{s=1}^{T-k}\frac{1}{(d/T^2+\lambda_{s,T-k}^{-1})^2}\right\},$$

where

$$g'_{k,T} = \left(\sum_{s=1}^{T-k}\frac{d^*/T^2+\lambda_{s,T-k}^{-1}}{d/T^2+\lambda_{s,T-k}^{-1}}w_s^2\right)^{-2}\left\{2\left(\sum_{s=1}^{T-k}\frac{d^*/T^2+\lambda_{s,T-k}^{-1}}{(d/T^2+\lambda_{s,T-k}^{-1})^3}w_s^2\right)\right.$$
$$\times\left(\sum_{s=1}^{T-k}\frac{d^*/T^2+\lambda_{s,T-k}^{-1}}{d/T^2+\lambda_{s,T-k}^{-1}}w_s^2\right)$$
$$\left.-\left(\sum_{s=1}^{T-k}\frac{d^*/T^2+\lambda_{s,T-k}^{-1}}{(d/T^2+\lambda_{s,T-k}^{-1})^2}w_s^2\right)^2\right\}.$$

The leading term, T^{-2} or T^{-4}, serves as a normalization factor in deriving the asymptotics for the processes, $l'_{k,T}(d)$ and $l''_{k,T}(d)$.

Now it follows easily that $E(l'_{k,T}(d^*)) = 0$. By contrast, the score of the profile likelihood is

$$\frac{1}{2T^2}\left\{(T-k)g_{k,T} - \sum_{s=1}^{T}\frac{1}{d/T^2+\lambda_{s,T}^{-1}}\right\},$$

where the deterministic term, $\lambda_{s,T-k}^{-1}$, is replaced by $\lambda_{s,T}^{-1}$, rendering expectation of the score nonzero.

3. ASYMPTOTIC RESULTS

We shall present two sets of asymptotic properties, consistency and the limiting distribution, for the local maximizer of the marginal likelihood.

The local maximizer is defined as the smallest positive root of the score. The definition suggests that the limiting behavior of the local maximizer be jointly determined by the limiting score and its limiting derivative. It is thus important to obtain the asymptotics for the score and its derivative. The following theorem describes the joint weak convergence of the score and its derivative as random functions on $C^2[0,\infty)$.

THEOREM 1. *Suppose* $0 \le d^*, d < \infty$, *then*

$$(l'_{k,T}(d), l''_{k,T}(d)) \Rightarrow (l'_k(d), l''_k(d)) \quad \textit{on } C^2[0,\infty),$$

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

where $k = 1,2$,

$$l_k'(d) = \frac{1}{2}\left\{\sum_{s=1}^{\infty}\frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)}\,w_s^2 - \sum_{s=1}^{\infty}\lambda_{s,k}(d)\right\},$$

and

$$l_k''(d) = \frac{1}{2}\left\{\left(\sum_{s=1}^{\infty}\frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)}\,w_s^2\right)^2 - 2\sum_{s=1}^{\infty}\frac{\lambda_{s,k}(d)^3}{\lambda_{s,k}(d^*)}\,w_s^2 + \sum_{s=1}^{\infty}\lambda_{s,k}(d)^2\right\}$$

with $w_s \sim \text{nid}(0,1)$, *and for* $c \geq 0$

$$\lambda_{s,1}(c) = \frac{1}{s^2\pi^2 + c},$$

$$\lambda_{2s-1,2}(c) = \frac{1}{4s^2\pi^2 + c}, \quad \textit{and} \quad \lambda_{2s,2}(c) = \frac{1}{\phi_s^2 + c},$$

where $\phi_s \in (2s\pi,(2s+1)\pi)$ *such that* $\tan\phi/2 = \phi/2$.

The result makes it clear how the conventional Taylor series approximation fails to derive the limiting distribution of the local maximizer. The derivative of the score is found to be random in the limit, whereas the standard limit distribution argument typically requires it to be asymptotically constant. In this nonstandard case, one may try to use other higher order terms in the Taylor series expansion. Unfortunately, any higher order expansion also proves to have a random limit as that of the second order.

By applying the same argument in proving Theorem 1, we can give a similar result for the local level model with a known start-up of the random walk component. The limit score process has the form

$$\frac{1}{2}\left\{\sum_{s=1}^{\infty}\frac{\lambda_s(d)^2}{\lambda_s(d^*)}\,w_s^2 - \sum_{s=1}^{\infty}\lambda_s(d)\right\},$$

where $\lambda_s(c) = 4/((2s-1)^2\pi^2 + 4c)$. The score derivative can be derived accordingly. Also note that when $k = 1$, the limiting score process is the same as that for the differenced local level model derived in Shephard (1993b).

To allow the zero boundary solution, our optimization problem invokes the Kuhn–Tucker conditions. Detailedly, $dl_{k,T}'(d) = 0$ and $l_{k,T}'(d) + dl_{k,T}''(d) \leq 0$, respectively, are the first- and the second-order conditions of the constrained maximization for small samples. Later, the conditions will be employed to pin down the local maximizer. The following result gives the joint limit of the conditions.

COROLLARY. *Suppose* $0 \leq d^*, d < \infty$,

$$(dl_{k,T}'(d),\, l_{k,T}'(d) + dl_{k,T}''(d)) \Rightarrow (dl_k'(d),\, l_k'(d) + dl_k''(d)) \quad \textit{on } C^2[0,\infty),$$

where $k = 1,2$.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

It is now possible to determine the convergence rate of the local maximizer estimator using the score process. The estimator is superconsistent in the sense that its convergence rate is faster than $\sqrt{T}$. Superconsistency result for estimators with nonstationary processes are often encountered. The result is stated in the next theorem.

THEOREM 2. *For arbitrary $\eta > 0$, there exists a $d_{\max} > 0$ such that*

$$\lim_{T\to\infty} \Pr(l'_{k,T}(d_{\max}) \geq 0 < \eta$$

where $k = 1,2$.

The result says that the score process can not be always greater than or equal to zero. In other words, the score process will cross zero sooner or later at some $d > 0$ after found positive at $d = 0$. Thus there exists a local maximizer in the interval, $[0, d_{\max}/T^2)$, with a high probability, i.e., $\hat{q} = O_p(1/T^2)$. The result extends Theorem 3 of Shephard (1993b) in two ways: to the case where the true parameter is near zero and to the dynamic regression model with a time trend. Sargan and Bhargava (1983) and Tanaka and Satchell (1989), on the other hand, showed that the ML estimator for a noninvertible MA(1) process is T-consistent.

Another version of Theorem 2 concerns the limit score process:

$$\Pr(l'_k(d) < 0 \quad \text{for some } d \geq 0) = 1, \quad \text{or} \quad \Pr(l'_k(d) \geq 0 \quad \text{for all } d \geq 0) = 0. \tag{7}$$

These assure that the limit score has a local root. The root can be a single point or a set of points in $R^+ = [0,\infty)$. A further analysis, however, suggests that the latter is unlikely. To see it, note that

$$\begin{aligned}
&(l'_k(d) < 0 \quad \text{for some } d \geq 0) \\
&\quad = (dl'_k(d) = 0 \quad \text{for some } d \geq 0) \\
&\quad = (dl'_k(d) = 0,\ l'_k(d) + dl''_k(d) < 0 \quad \text{for some } d \geq 0) \\
&\qquad + (dl'_k(d) = l'_k(d) + dl''_k(d) = 0 \quad \text{for some } d \geq 0).
\end{aligned}$$

The first equivalence relation holds because a score process taking a negative value somewhere in R^+ must be the case that it is negative when evaluated at the origin, or that it crosses zero somewhere in R^+. Either event implies $dl'_k(d) = 0$. The second term in the right-hand side of the second equality is the event that the local root is a set of points and is shown in the Appendix to be of zero probability. Therefore

$$\Pr(dl'_k(d) = 0,\ l'_k(d) + dl''_k(d) < 0 \quad \text{for some } d \geq 0) = 1.$$

That is, the limit local root is always a single point. Because it must be contained in a certain interval, in practice this property makes applicable a grid search that is to locate the interval.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

The event just defined indeed includes all the possible local maximizers, zero or greater than zero. Using the limit Kuhn–Tucker conditions, we see this because

$$\begin{cases} c = 0 & \text{iff } 0l_k'(0) = 0,\ l_k'(0) + 0l_k''(0) = l_k'(0) < 0, \quad \text{and} \\ c > 0 & \text{iff } cl_k'(c) = 0,\ l_k'(c) + cl_k''(c) = cl_k''(c) < 0, \end{cases} \tag{8}$$

where c denotes a local root. Among these maximizers it is the smallest one that is of major interest. We are thus in a position to define the limit local maximizer estimator:

$$\hat{d}_k = \inf_{d \geq 0} (dl_k'(d) = 0,\ l_k'(d) + dl_k''(d) < 0).$$

Correspondingly, the empirical local maximizer estimator is defined as

$$\hat{d}_{k,T} = \begin{cases} \inf_{d \geq 0} (dl_{k,T}'(d) = 0,\ l_{k,T}'(d) + dl_{k,T}''(d) < 0) & \text{if } M_T \neq \varnothing \\ \inf_{d \geq 0} (dl_{k,T}'(d) = 0,\ l_{k,T}'(d) + dl_{k,T}''(d) \leq 0) & \text{if } M_T = \varnothing, \quad \text{and} \\ & M_T^0 \neq \varnothing, \end{cases}$$

where $M_T = \{(l_{k,T}', l_{k,T}'') | dl_{k,T}'(d) = 0, l_{k,T}'(d) + dl_{k,T}''(d) < 0 \text{ for some } d \geq 0, l_{k,T}'(\cdot)$ and $l_{k,T}''(\cdot) \in C[0,\infty)\}$, and $M_T^0 = \{(l_{k,T}', l_{k,T}'') | dl_{k,T}'(d) = 0,\ l_{k,T}'(d) + d_{k,T}''(d) = 0$ for some $d \geq 0$, $l_{k,T}'(\cdot)$ and $l_{k,T}''(\cdot) \in C[0,\infty)\}$. The latter is an event of the empirical root being a set of points. But the event is asymptotically negligible because it is the sample version of the event $(dl_k'(d) = l_k'(d) + dl_k''(d) = 0$ for some $d \geq 0)$. By its definition, the sampling local maximizer is zero if $l_{k,T}'(0) < 0$. Otherwise, it is the smallest positive root of the score process. The definition is parallel to that of the limit local maximizer.

It should be emphasized that $\hat{d}_k$ or $\hat{d}_{k,T}$ is a random variable. It is a continuous mapping from the space $C^2[0,\infty)$ to R^+. Having established the notations, we are able to give the distribution result for the local maximizer.

THEOREM 3. *Let* $\hat{q}_k = \hat{d}_{k,T}/T^2$, *then*

$$T^2\hat{q}_k \underset{d}{\rightarrow} \hat{d}_k,$$

where $k = 1,2$.

The limit distribution has a number of features. First, it is continuous except at zero. This can be seen from (8) simply by recognizing that for $c > 0$, $\Pr(c > 0) = \Pr(l_k'(c) = 0)$. We then have the continuity of d_k at any point greater than 0 as a result of the continuity of $l_k'(c)$. Second, the limit distribution has a mass point at zero. This is because by (8) again,

$$\begin{aligned} \lim_{T\to\infty} \Pr(T^2\hat{q}_k = 0) &= \lim_{T\to\infty} \Pr(\hat{d}_k = 0) \\ &= \lim_{T\to\infty} \Pr(l_{k,T}'(0) \leq 0) \\ &= \Pr(l_k'(0) \leq 0). \end{aligned}$$

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

The formula is an alternative of computing the probability of a jump at zero to the Fredholm determinant approach. The latter approach, developed in Nabeya and Tanaka (1988), analytically computes the probability with characteristic function of the limit score process. We can understand now how the deterministic term in the score affects the probability that the local maximizer is estimated to be zero. It is evident that

$$\Pr(l_k'(0) \le 0) = \Pr\left(\left(\sum_{s=1}^{\infty} \frac{\lambda_{s,k}(0)^2}{\lambda_{s,k}(d^*)} w_s^2 - \sum_{s=1}^{\infty} \lambda_{s,k}(0)\right) \le 0\right)$$
$$< \Pr\left(\left(\sum_{s=1}^{\infty} \frac{\lambda_{s,k}(0)^2}{\lambda_{s,k}(d^*)} w_s^2 - \sum_{s=1}^{\infty} \lambda_s(0)\right) \le 0\right)$$

as a result of $\sum_{s=1}^{\infty}\lambda_{s,k}(0) < \sum_{s=1}^{\infty}\lambda_s(0) = \frac{1}{2}$. The preceding inequality states that the profile likelihood is more likely to produce zero estimates for the local root than the marginal likelihood. This justifies the use of the marginal likelihood in contexts like the one given here.

Likewise, the limit distribution for the local level model with a known start-up has a similar representation. Although we have been focusing on deriving the limit distribution of the local maximizer for the MML estimator, the limit distribution for the ML estimator can be derived without difficulty. Specifically, it is found to be

$$\inf_{d\ge 0}(dl_k^{p\prime}(d) = 0,\ l_k^{p\prime}(d) + dl_k^{p\prime\prime}(d) < 0),$$

where

$$l_k^{p\prime}(d) = \frac{1}{2}\left\{\sum_{s=1}^{\infty} \frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)} w_s^2 - \sum_{s=1}^{\infty} \lambda_s(d)\right\},$$

and

$$l_k^{p\prime\prime}(d) = \frac{1}{2}\left\{\left(\sum_{s=1}^{\infty} \frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)} w_s^2\right)^2 - 2\sum_{s=1}^{\infty} \frac{\lambda_{s,k}(d)^3}{\lambda_{s,k}(d^*)} w_s^2 + \sum_{s=1}^{\infty} \lambda_s(d)^2\right\},$$

$w_s \sim \text{nid}(0,1)$.

As expected, it only differs from that for the MML estimator, $\hat{d}_k$, by the deterministic term in the score and its derivative processes.

4. SIMULATION RESULTS

This section is devoted to evaluating whether the limiting distribution derived in Section 3 provides a useful approximation to the finite-sample distribution. Shephard (1993b) made an attempt to approximate the distribution of the ML estimator for the local level model. His technique yields a good approximation yet is different from our asymptotic approximation. As another distinction, his ap-

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

proach focuses on the distribution of the global maximizer, whereas here attention is paid to that of the local maximizer.

The simulation is based on Theorem 3, whose representation, however, is not a closed form solution to the local maximizer estimator. This suggests that to unveil the limit distribution of $T^2\hat{q}$ one virtually has to resort to the numerical approach. We adopt the following procedures. We draw an independent series of 5,000 terms from the standard normal as an approximation to the infinite series $l_k'(d)$. For any particular replicate like this, $\bar{d}_k$ takes a zero value if the sign of the truncated series evaluated at zero is negative. Otherwise the calculation proceeds until the smallest positive root for $l_k'(d) = 0$ is found. The term $\bar{d}_k$ is then equal to this root for the draw. To make the root finding more accurate, in addition to a stepwise search, a binary search is also called to assure computation precision when the truncated process makes its first zero-crossing.[2] The same root-finding procedures are also applied to locate the zero-crossing for the sampling distributions. All the results presented subsequently were performed on UNIX with FORTRAN and based on 20,000 replications.

The local level model, equations (3) and (1), are examined in order, i.e.,

$$y_t = \gamma_t + u_t, \qquad y_t = \alpha + \gamma_t + u_t, \quad \text{and} \quad y_t = \alpha + \beta t + \gamma_t + u_t.$$

The simulation is mainly to compare the asymptotic distribution of $T^2\hat{q}$ with the sampling distributions for different sample sizes when d^* is zero or near zero. However, it is important to have a likelihood that sacrifices as little information as possible. When there are more regression parameters involved, the distribution of the local maximizer is expected to be more dispersed. This is a reflection of information loss due to additional parameters. The simulation displayed subsequently is not at odds with the expectation.

Figure 1 plots the asymptotic distribution of $T^2\hat{q}$ against the sampling ones for the local level model when $d^* = 0$. Four different sample sizes are chosen, 10, 20, 50, and 100. Table 1 gives the numerical aspect of the simulation. As it illustrates, the limit distribution is in accordance with the sampling distributions for samples of moderate size, 50 and 100. As will be seen subsequently, because there is no regression parameter involved, the limit distribution in this model when the sample is of these sizes gives the best approximation among the three models under study. A long tail and a mass point at zero appear to be common features to both the limit and sampling distributions. Figure 2 and Table 2 demonstrate this similar pattern when the true parameter d^* shifts against zero. Simulations here were performed for selective values of $d^* = .25, 1.25, 2.5, 12.5$ and for sample size $T = 50, 100$. But the approximation is good only for those $d^* \leq 25$. This is because the local-to-zero parameterization is intended to cover the cases when d^* is near zero. We anticipate that the normal limit approximation would outperform for the cases when d^* is greater than 25. The pictures indicate that there exists a positive probability of the local maximizer being zero that slowly diminishes as d^* moves to the right. The probability of a jump at zero for finite samples is seen to be accurately approximated with the formula $\Pr(l_k'(0) \leq 0)$. In the case that $d^* = 0$,

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

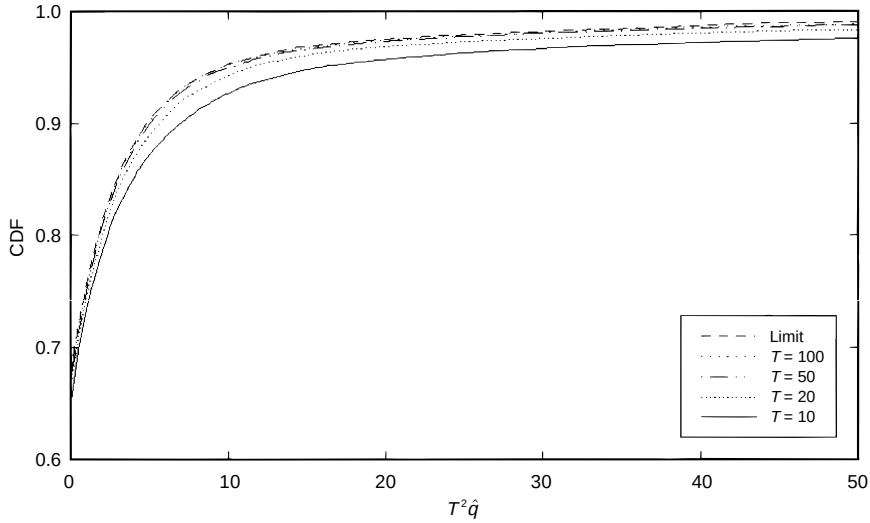


FIGURE 1. The distribution of $T^2\hat{q}$ when true q is on the boundary (model with stochastic trend component).

the limiting probability that $\hat{q} = 0$ is also found to totally agree with the analytical one derived from the Fredholm determinant approach. The numerical results basically resemble those reported in Shephard and Harvey (1990) for the local level model with a known fixed start-up.

We turn to the simulation for models with an unknown intercept and a stochastic trend component. Along with those for equation (1), the simulation setup here follows as before, except in the boundary case where only sample sizes $T = 50$

TABLE 1. Limit and empirical distributions of $T^2\hat{q}$ when q^* is zero: Local level model with a known start-up

	p					
T	.01	.05	.10	.15	.20	$\Pr(\hat{q} = 0.)$
10	219.02	15.52	6.98	3.95	2.30	.655
20	79.31	11.19	5.96	3.28	1.99	.668
50	58.76	9.98	5.33	3.03	1.81	.674
100	54.00	9.33	4.86	2.98	1.79	.677
∞	44.85	9.43	4.83	2.92	1.71	.678

Notes: 20,000 replications are generated for each simulation. The limit case is approximated by a draw of $T = 5{,}000$ in each replicate.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

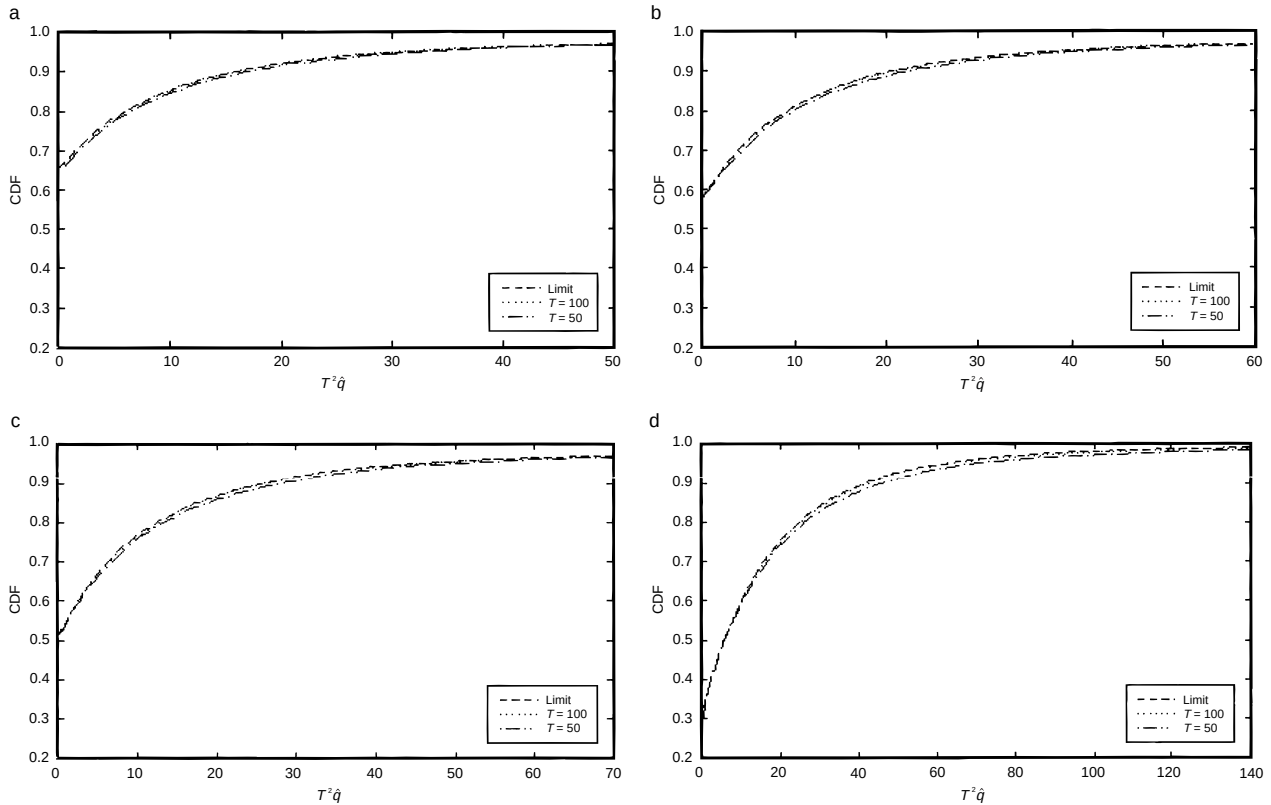


FIGURE 2. The distribution of $T^2\hat{q}$ when true q is near the boundary (model with stochastic trend component): (a) deviation = 0.25; (b) deviation = 1.25; (c) deviation = 2.5; (d) deviation = 12.5.

TABLE 2. Limit and empirical distributions of $T^2\hat{q}$ when q^* is near zero: Local level model with a known start-up

		p					
d^*	T	.01	.05	.10	.15	.20	$\Pr(\hat{q} = 0.)$
.25	50	56.60	11.44	6.02	3.71	2.32	.649
	100	49.19	10.57	5.49	3.43	2.17	.651
	∞	48.18	10.82	5.60	3.48	2.18	.654
1.25	50	74.14	17.66	9.60	6.24	4.24	.574
	100	62.18	16.02	8.74	5.73	4.02	.577
	∞	59.81	16.42	8.83	5.84	3.95	.577
2.5	50	92.14	24.68	13.83	9.33	6.64	.507
	100	76.70	22.78	12.79	8.69	6.30	.511
	∞	70.65	22.39	12.88	8.66	6.19	.511
12.5	50	179.21	70.77	47.75	33.96	26.13	.301
	100	149.68	63.91	43.73	31.76	24.91	.301
	∞	134.93	63.05	42.96	31.25	24.56	.298

Notes: 20,000 replications are generated for each simulation. The limit case is approximated by a draw of $T = 5{,}000$ in each replicate.

and 100 are considered for the sampling distributions. Again, a long tail and a mass point at zero are features of Figures 3 and 4. Both the limit and the empirical distributions now assume larger values for higher quantiles than those reported before. This implies that both the limit and the sampling distributions are characterized by a longer right-hand tail, and therefore the local maximizer estimates are more scattered. Further, it appears that the speed of the convergence to the limit distribution is not as fast as before. But the limit approximation is still good for sample size $T = 100$. It calls for more sample size before the limit approximation is exact. On the other hand, we observe that the probability that the estimate is zero is more persistent for d^*'s near zero, whereas it takes a slightly smaller value when $d^* = 0$. The findings are consistent with the simulation results of Shephard (1993a). The results presented in Figures 3 and 4 and Table 3 could shed light on the first-differenced local level model, which, as noted before, has the same likelihood as that of the model of concern here.

The simulation results for models with an intercept, a time trend, and a stochastic trend component are similar as before. Overall, the additional unknown regression coefficient causes a longer tail and a zero mass point with a smaller probability. The numerical results are summarized in Table 4. Plots for the limit and empirical distributions are not given, for they are not qualitatively different from the previous ones.

Figure 5 shows how introducing regression parameters into the model affects the limit distribution. In that figure the shapes of the limit distributions for models

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

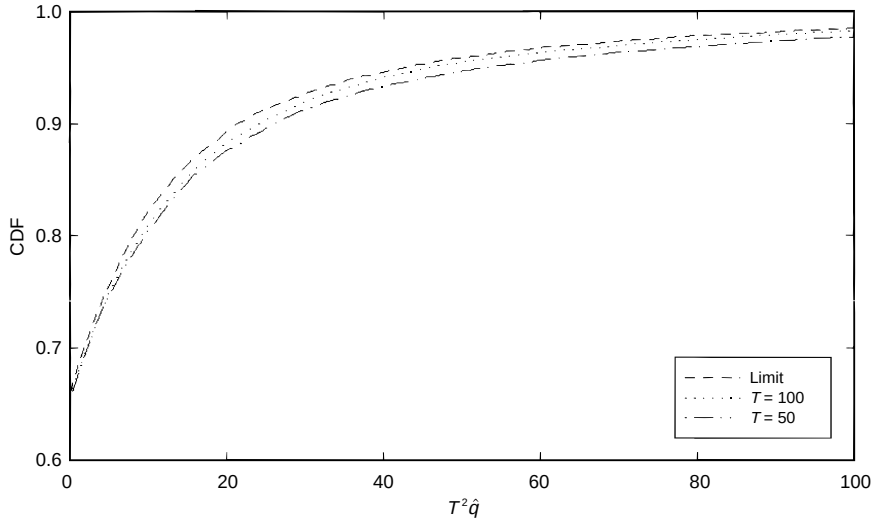


FIGURE 3. The distribution of $T^2\hat{q}$ when true q is zero (model with intercept and stochastic trend component).

TABLE 3. Limit and empirical distributions of $T^2\hat{q}$ when q^* is zero and near zero: Model with intercept and stochastic trend components

		p					
d^*	T	.01	.05	.10	.15	.20	$\Pr(\hat{q} = 0.)$
.0	50	174.11	52.38	25.92	15.14	9.47	.654
	100	133.21	45.28	24.03	14.65	9.01	.657
	∞	121.30	42.59	21.42	13.05	7.96	.661
.25	50	175.90	53.47	26.73	15.85	10.02	.647
	100	135.16	46.40	24.79	15.25	9.49	.650
	∞	122.87	43.54	22.21	13.69	8.45	.655
1.25	50	183.03	58.08	30.14	18.53	12.08	.623
	100	141.76	50.59	27.96	17.61	11.38	.626
	∞	128.81	47.78	25.11	16.02	10.32	.630
2.5	50	194.72	64.68	34.51	21.87	14.64	.597
	100	149.68	55.71	31.81	20.58	13.94	.597
	∞	138.58	52.73	28.87	18.96	12.71	.602
12.5	50	269.77	108.29	66.55	47.06	35.05	.448
	100	204.30	93.59	60.79	43.61	33.58	.447
	∞	189.13	88.10	56.18	41.16	30.93	.450

Notes: 20,000 replications are generated for each simulation. The limit case is approximated by a draw of $T = 5{,}000$ in each replicate.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

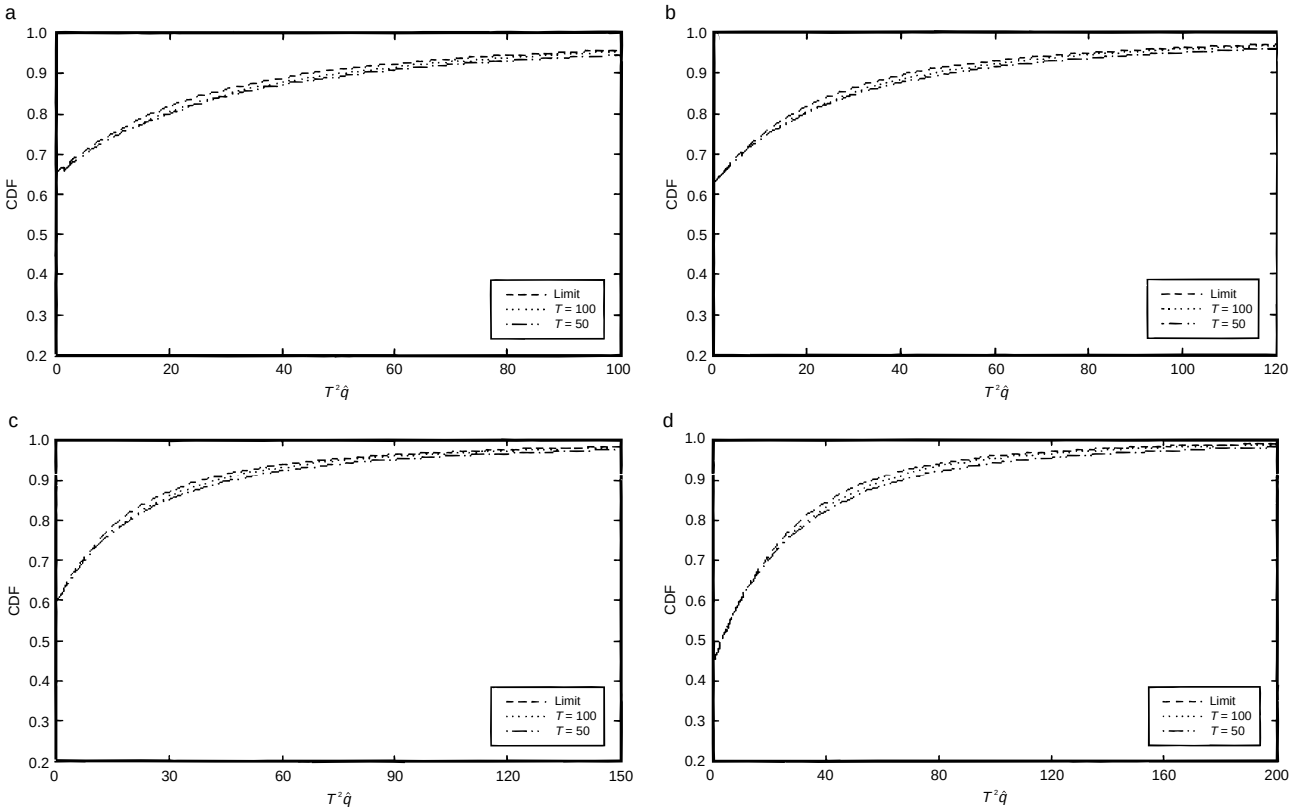


FIGURE 4. The distribution of $T^2\hat{q}$ when true q is near zero (model with intercept and stochastic trend component): (a) $d^* = 0.25$; (b) $d^* = 1.25$; (c) $d^* = 2.5$; (d) $d^* = 12.5$.

TABLE 4. Limit and empirical distributions of $T^2\hat{q}$ when q^* is zero and near zero: Model with intercept, time trend, and stochastic trend components

		p					
d^*	T	.01	.05	.10	.15	.20	$\Pr(\hat{q} = 0.)$
.0	50	441.28	167.99	95.20	61.94	41.05	.618
	100	333.92	142.87	86.13	56.98	39.31	.619
	∞	291.06	130.12	77.80	51.13	34.04	.625
.25	50	442.22	168.81	95.86	62.53	41.57	.617
	100	335.28	143.85	86.95	57.50	39.81	.617
	∞	291.98	130.79	78.35	51.64	34.49	.623
1.25	50	450.32	172.63	98.57	64.80	43.47	.610
	100	341.16	148.02	89.78	59.87	41.80	.610
	∞	295.27	133.95	80.80	53.61	36.14	.612
2.5	50	459.99	177.37	101.93	67.83	45.82	.602
	100	346.20	152.12	93.09	62.56	44.02	.601
	∞	298.51	137.20	83.84	56.05	38.26	.606
12.5	50	517.26	212.87	130.87	90.16	65.08	.539
	100	389.17	184.58	118.17	84.07	62.34	.537
	∞	337.29	167.15	107.81	76.08	56.11	.537

Notes: 20,000 replications are generated for each simulation. The limit case is approximated by a draw of $T = 5{,}000$ in each replicate.

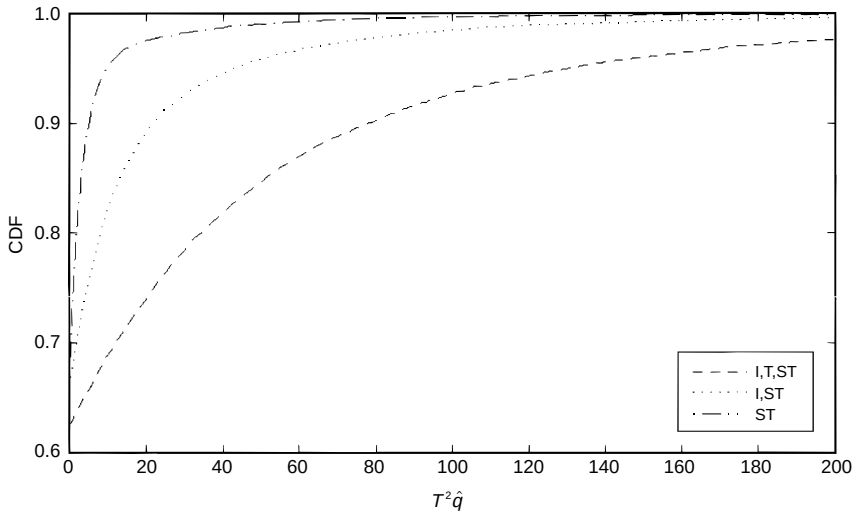


FIGURE 5. The comparison of limiting distribution when true q is near zero (I, intercept; T, time trend; ST, stochastic trend).

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

of interest are together illustrated for the case $d^* = 0$. The figure substantiates that the additional nuisance parameter gives rise to more dispersion.

Given that the limit approximation is reasonably good, the asymptotic quantiles reported for $d^* = 0$ can therefore be used as critical values in testing the null stationarity that $H_0 : q = 0$ against $H_\alpha : q > 0$. In other words, it is likely to accept the null at the α% level if the local maximizer estimate is smaller than the $(1 - \alpha)$th quantiles, $\hat{d}_k(\alpha)$, such that

$$\Pr(\hat{d}_k > \hat{d}_k(\alpha)) = \alpha.$$

For example, when equation (1) is estimated, for $T = 100$, the stationarity null is accepted at the 5% level if $(100^2)\hat{q}_k < 142.87$. The new test is different from those derived by the Lagrange multiplier principle such as those of Kwiatkowski et al. (1992) and Schmidt and Phillips (1992). The superiority of these tests remains to be seen.

5. CONCLUDING REMARKS

This paper mainly examined the asymptotic properties of the local maximizer for models with a stochastic trend component. Our results complement those of D&D (1996) for the case that the moving average parameter is close to 1 in the MA(1) model. A couple of interesting problems arise that invite further research. In theory, it may be useful to derive corresponding asymptotics for the global maximizer using the results here. This type of weak convergence properties is essential to establishing new Wald-type or LR-type tests for the null of stationarity. Moreover, performance comparisons between these new tests and the LM-type tests appear to be another area of research worth pursuing in the future.

NOTES

1. Generally, the derivative of $L_k^p(q; y)$ can be written as

$$\frac{dL_k^p(q;y)}{dq} = \frac{dL_k(q, \hat{\delta}(q), \hat{\sigma}^2(q); y)}{dq} = \frac{dL_k}{dq} + \frac{dL_k}{d\delta}\frac{d\hat{\delta}(q)}{dq} + \frac{dL_k}{d\sigma^2}\frac{d\hat{\sigma}^2(q)}{dq},$$

where L_k is the original log-likelihood. Taking the expectation, it can be found that the last two terms obviously are not identically zero, though the first term is. Shortly, we will calculate the score mean of the profile likelihood, along with that of the marginal likelihood introduced later.

2. Given that the root is bounded (Theorem 2), the binary search in general enables a faster root finding if a right-hand starting guess value is chosen appropriately large. However it does not always lead to the local maximizer solution of interest as the limiting score could have multiple roots, including the global maximizer. The stepwise search, followed by the binary search, is thereby first adopted to locate the interval containing the root.

REFERENCES

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: Harvard University Press.

Anderson, T.W. & A. Takemura (1986) Why do noninvertible estimated moving averages occur? *Journal of Time Series Analysis* 7, 235–254.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

Billingsley, P. (1968) *Convergence of Probability Measures*. New York: Wiley.
Cooley, F.T. & E.C. Prescott (1976) Estimation in the presence of stochastic parameter variation. *Econometrica* 44, 167–184.
Davis, R.A. & W.T.M. Dunsmuir (1996) Maximum likelihood estimation for MA(1) processes with a root on or near the unit circle. *Econometric Theory* 12, 1–29.
Dhrymes, P.J. (1984) *Mathematics for Econometrics*. New York: Springer-Verlag.
Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, U.K.: Cambridge University Press.
Harvey, A.C., B. Henry, S. Peters, & S. Wren-Lewis (1986) Stochastic trends in dynamic regression models: An application to the employment-output equation. *Economic Journal* 96, 975–985.
Harvey, A.C. & J.H. Stock (1988) Continuous time autoregressive models with common stochastic trends. *Journal of Economic Dynamics and Control* 12, 365–384.
Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, & Y.C. Shin (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 54, 159–178.
McCullagh, P. & J.A. Nelder (1989) *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.
Muth, J.F. (1960) Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55, 299–305.
Nabeya, S. & K. Tanaka (1988) Asymptotic theory of a test for the constancy of regression coefficients against the random walk alternative. *Annals of Statistics* 16, 218–235.
Nyblom, J. (1986) Testing for deterministic linear trend in time series. *Journal of the American Statistical Association* 81, 545–549.
Nyblom, J. & T. Mäkeläinen (1983) Comparisons of tests for the presence of random walk coefficients in a simple linear model. *Journal of the American Statistical Association* 78, 856–864.
Patterson, H.D. & R. Thompson (1971) Recovery of interblock information when block sizes are unequal. *Biometrika* 58, 545–554.
Rao, C.R. (1973) *Linear Statistical Inference and Its Applications*. New York: Wiley.
Saikkonen, P. & R. Luukkonen (1993) Testing for a moving average unit root in autoregressive integrated moving average models. *Journal of the American Statistical Association* 88, 596–601.
Sargan, J.D. & A. Bhargava (1983) Maximum likelihood estimation of regression models with first-order moving average errors when the root lies on the unit circle. *Econometrica* 51, 799–820.
Schmidt, P. & P.C.B. Phillips (1992) LM test for a unit root in the presence of deterministic trends. *Oxford Bulletin of Economics and Statistics* 54, 257–287.
Shephard, N. (1993a) Maximum likelihood estimation of regression models with stochastic trend components. *Journal of the American Statistical Association* 88, 590–595.
Shephard, N. (1993b) Distribution of the ML Estimator of a MA(1) and a local level model. *Econometric Theory* 9, 377–401.
Shephard, N. & A.C. Harvey (1990) On the probability of estimating a deterministic component in the local level model. *Journal of Time Series Analysis* 11, 339–347.
Tanaka, K. & S.E. Satchell (1989) Asymptotic properties of the maximum likelihood and nonlinear least squares estimator for noninvertible moving average models. *Econometric Theory* 5, 333–353.

APPENDIX: MATHEMATICAL PROOFS

Lemmas

LEMMA 1.

$$\sum_{s=1}^{T-k} \log(1 + q\lambda_{s,T-k}) = \log|\Omega(q)| + \log|X'\Omega(q)^{-1}X| + 2\log|G|,$$

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

where G is a $k \times k$ transformation matrix such that $U = XG'$, where U is a $T \times k$ orthogonal matrix.

Proof. Let $V = (W, U)$, where W is an orthogonal matrix as defined before. Thus $VV' = V'V = I_T$. Now

$$\begin{aligned}
|\Omega(q)| &= |V'\Omega(q)V| \\
&= |W'\Omega(q)W||U'(\Omega(q) - \Omega(q)W(W'\Omega(q)^{-1}W)^{-1}W'\Omega(q))U| \\
&= |W'\Omega(q)W||U'U(U'\Omega(q)^{-1}U)^{-1}U'U| \\
&= |W'\Omega(q)W|/|U'\Omega(q)^{-1}U| \\
&= |W'\Omega(q)W|/|GX'\Omega(q)^{-1}XG'| \\
&= |W'\Omega(q)W|/|X'\Omega(q)^{-1}X||G|^2.
\end{aligned}$$

The second equivalence is based on a result in Dhrymes (1984, p. 37). The third equivalence exploits the relation in Rao (1973, p. 77) that $\Sigma - \Sigma M(M'\Sigma^{-1}M)^{-1}M'\Sigma = N(N'\Sigma^{-1}N)^{-1}N'$ where Σ is any positive definite matrix, M, N are $T \times (T-k)$ and $T \times k$ matrices such that if $O = (M, N)$ then $OO' = O'O = I_T$. Using the fact that $|W'\Omega(q)W| = \prod_{s=1}^{T-k}(1 + q\lambda_{s,T-k})$ and taking the logarithm on both sides, we have the result as required.

LEMMA 2.

$$\Pr(dl_k'(d) = dl_k''(d) + l_k'(d) = 0 \quad \textit{for some } d \geq 0) = 0.$$

Proof. The proof follows the style of Proposition A.1 in D&D. Conditional on $(w_2, w_3, \ldots)$, there exists $d = d(w_1)$ such that $l_k'(d) = l_k''(d) = 0$. This is an equation system with two unknowns d and w_1. We will show that the solution d is countable. Solving $l_k'(d) = 0$ for w_1, and plugging it into $l_k''(d) = 0$, it yields

$$\begin{aligned}
l_k''(d) = \frac{1}{2}\Bigg\{ &\left[\frac{\lambda_{s,1}(d^*)}{\lambda_{s,1}(d)} \left(\sum_{s=1}^{\infty} \lambda_{s,k}(d) - \sum_{s=2}^{\infty} \frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)} w_s^2 \right) + \sum_{s=2}^{\infty} \frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)} w_s^2 \right]^2 \\
&- 2\left[\frac{\lambda_{s,1}(d^*)}{\lambda_{s,1}(d)} \left(\sum_{s=1}^{\infty} \lambda_{s,k}(d) - \sum_{s=2}^{\infty} \frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)} w_s^2 \right) + \sum_{s=2}^{\infty} \frac{\lambda_{s,k}(d)^3}{\lambda_{s,k}(d^*)} w_s^2 \right] \\
&+ \sum_{s=1}^{\infty} \lambda_{s,k}(d)^2 \Bigg\} = 0.
\end{aligned}$$

Now extending the permissible space to allow $d \in (-\pi^2, \infty)$ (for $k = 1$), or $(-2\pi^2, \infty)$ (for $k = 2$), it is easy to find that $l_k''(d) \to \infty$ as $d \to -\pi^2$ (for $k = 1$), or as $d \to -2\pi^2$ (for $k = 2$), and $l_k''(d) \to -\infty$ as $d \to \infty$. The term $l_k''(d)$ is thus not always zero, and the zeros of $l_k''(d)$ are isolated. Thus $d = d(w_1)$ is countable. Because w_1 is solely determined by $l_k'(d) = 0$ conditional on $(w_2, w_3, \ldots)$, it also implies that w_1 is countable.

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

Let $S = S(w_2, w_3, \ldots)$ denote the set of w_1 for which $l_k'(d) = l_k''(d) = 0$ for some $d \geq 0$. Thus we deduce that

$$\Pr(w_1 \in S | w_2, w_3, \ldots) = \Pr(dl_k'(d) = dl_k''(d) + l_k'(d) = 0 \text{ for some } d \geq 0 | w_2, w_3, \ldots) = 0.$$

The result follows. ■

Proof of Theorems

Proof of Theorem 1. We first establish the finite-dimensional convergence. For any $s > 0$, and $d \in [0, N]$, where $N < \infty$,

$$\frac{1}{T^2} a_{s,T} = \frac{1}{T^2} \frac{d^*/T^2 + \lambda_{s,T-k}^{-1}}{(d/T^2 + \lambda_{s,T-k}^{-1})^2} w_s^2 \Rightarrow a_s = \frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)} w_s^2 \quad \text{as } T \to \infty.$$

Immediate on applying the Slutsky theorem and the continuous mapping theorem (CMT), if m is an integer,

$$\frac{1}{T^2} \sum_{s=1}^{m} a_{s,T} \Rightarrow \sum_{s=1}^{m} a_s \quad \text{on } C[0, N].$$

To show that the convergence result holds as $m \to \infty$, note that for all $d \in [0, N]$,

$$\limsup_{T \to \infty} E\left(\sup_{d \geq 0} \frac{1}{T^2} \sum_{s=m+1}^{n_T} a_{s,T} \right) \leq \sum_{s=m+1}^{\infty} \frac{\lambda_{s,k}(0)^2}{\lambda_{s,k}(d^*)} + o(1) \to 0,$$

where n_T is a sequence of integers such that $n_T \to \infty$, $n_T/T \to 0$, and $n_T^2/T \to \infty$ as $T \to \infty$. Then following from Theorem 4.2 in Billingsley (1968, p. 25),

$$\frac{1}{T^2} \sum_{s=1}^{m} a_{s,T} \Rightarrow \sum_{s=1}^{\infty} a_s \quad \text{on } C[0, N]. \tag{A.1}$$

Next, note that for all $d \in [0, N]$,

$$\frac{1}{T^2} c_{s,T} = \frac{1}{T^2} \left(\frac{d}{T^2} + \frac{1}{\lambda_{s,T-k}} \right)^{-1} \to c_s = \lambda_{s,k}(d).$$

Because $c_{s,T} \leq \lambda_{s,1}(0)$, $\forall s > 0$, and $\sum_{s=1}^{\infty} \lambda_{s,1}(0) = \frac{1}{6} < \infty$, we conclude that for all $d \in [0, N]$,

$$\frac{1}{T^2} \sum_{s=1}^{T-k} c_{s,T} \to \sum_{s=1}^{\infty} c_s. \tag{A.2}$$

Further, we show that on $C[0, N]$,

$$\frac{1}{T-k} \sum_{s=1}^{T-k} b_{s,T} = \sum_{s=1}^{T-k} \frac{d^*/T^2 + \lambda_{s,T-k}^{-1}}{d/T^2 + \lambda_{s,T-k}^{-1}} w_s^2 \to_p 1. \tag{A.3}$$

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

By Chebyshev's inequality, for any $\delta > 0$,

$$\Pr\left\{\left|\frac{1}{T-k}\sum_{s=1}^{T-k}\frac{d^*/T^2+\lambda_{s,T-k}^{-1}}{d/T^2+\lambda_{s,T-k}^{-1}}w_s^2-\frac{1}{T-k}\sum_{s=1}^{T-k}w_s^2\right|>\delta\right\}$$

$$<\frac{|d-d^*|}{T-k}\,\frac{E\left|\sum_{s=1}^{T-k}\frac{1}{T^2(d/T^2+\lambda_{s,T-k}^{-1})}w_s^2\right|}{\delta}$$

$$=\frac{(\text{const})\sum_{s=1}^{T-k}\lambda_{s,k}(d)}{T-k}+o(1)$$

$$\to 0,$$

giving

$$\frac{1}{T-k}\sum_{s=1}^{T-k}\frac{d^*/T^2+\lambda_{s,T-k}^{-1}}{d/T^2+\lambda_{s,T-k}^{-1}}w_s^2=\frac{1}{T-k}\sum_{s=1}^{T-k}w_s^2+o_p(1).$$

But $(1/T-k)\sum_{s=1}^{T-k}w_s^2\to_p 1$, so we have (A.3).

Using again the Slutsky theorem and the CMT, (A.1)–(A.3) give

$$l'_{k,T}(d)=\frac{1}{2T^2}\left\{(T-k)\frac{\sum_{s=1}^{T-k}a_{s,T}}{\sum_{s=1}^{T-k}b_{s,T}}-\sum_{s=1}^{T-k}c_{s,T}\right\}\Rightarrow l'_k(d)$$

$$=\frac{1}{2}\left\{\sum_{s=1}^{\infty}\frac{\lambda_{s,k}(d)^2}{\lambda_{s,k}(d^*)}w_s^2-\sum_{s=1}^{\infty}\lambda_{s,k}(d)\right\}$$

on $C[0,N]$. Similarly,

$$l''_{k,T}(d)\Rightarrow l''_k(d)\quad\text{on } C[0,N].$$

Together, we conclude the following joint convergence,

$$(l'_{k,T}(d),\,l''_{k,T}(d))\Rightarrow(l'_k(d),\,l''_k(d))\quad\text{on } C^2[0,N].$$

But because N is arbitrary, the joint convergence also holds for every $N>0$ and hence on $C^2[0,\infty)$ as asserted. The corollary is simply a consequence of applying the CMT and the Slutsky theorem. ■

Proof of Theorem 2. The proof is similar to that of Proposition 1 of Tanaka and Satchell (1989). A little calculation gives

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

$$E(l_k'(d_{\max})) = \frac{(d^* - d_{\max})}{2} \sum_{s=1}^{\infty} \lambda_{s,k}(d_{\max})^2, \quad \text{and} \quad \operatorname{Var}(l_k'(d_{\max})) = \frac{3}{4} \sum_{s=1}^{\infty} \frac{\lambda_{s,k}(d_{\max})^4}{\lambda_{s,k}(d^*)^2}.$$

By Chebyshev's inequality,

$$\begin{aligned}\Pr(l_k'(d_{\max}) \geq 0) &< \frac{\operatorname{Var}(l_k'(d_{\max}))}{[E(l_k'(d_{\max}))]^2} \\ &= 3\left(\sum_{s=1}^{\infty} \frac{\lambda_{s,k}(d_{\max})^4}{\lambda_{s,k}(d^*)^2}\right)\left(\sum_{s=1}^{\infty} \lambda_{s,k}(d_{\max})^2\right)^{-2} (d_{\max} - d^*)^{-2} \\ &\leq \frac{C}{(d_{\max} - d^*)^2},\end{aligned}$$

where C is a finite number such that $\forall d \geq 0$, $0 < 3(\sum_{s=1}^{\infty} \lambda_{s,k}(d)^4 \lambda_{s,k}(d^*)^{-2}) \times (\sum_{s=1}^{\infty} \lambda_{s,k}(d)^2)^{-2} \leq C$. Letting $d_{\max} = d^* + \sqrt{C/\eta}$ leads to the required inequality. ■

Proof of Theorem 3. Define the mapping $f: C^2[0,\infty) \to R^+ = [0,\infty)$ by

$$f(g_1, g_2) = \inf_{d \geq 0} \{dg_1(d) = 0, g_1(d) + dg_2(d) < 0, \quad \text{where } (g_1, g_2) \in C^2[0,\infty)\}.$$

Also let $H = \{(h, h') | dh(d) = 0, h(d) + dh'(d) < 0, \text{ for some } d \geq 0, h(\cdot) \text{ and } h'(\cdot) \in C[0,\infty)\}$. So $H \subset C^2[0,\infty)$.

We shall first show that f is continuous in H. Suppose $f(g) = \bar{d} > 0$ where $g = (h, h') \in H$ and let $g_t = (h_t, h_t') \to g = (h, h')$. By the definition of f, it must be the case that $h(\bar{d}) = 0$ and $h'(\bar{d}) < 0$. This suggests that we can find a neighborhood $(\bar{d} - \eta, \bar{d} + \eta)$ such that $h(\bar{d} + \eta) < 0 < h(\bar{d} - \eta)$ and $h'(z)\ \forall z \in (\bar{d} - \eta, \bar{d} + \eta)$. Similarly, because by construction $(h_t, h_t') \to (h, h')$, we should have that $h_t(\bar{d} + \eta) < 0 < h_t(\bar{d} - \eta)$ and $h_t'(z)\ \forall z \in (\bar{d} - \eta, \bar{d} + \eta)$ when t is sufficiently large. This implies that there exists some $d_t \in (\bar{d} - \eta, d + \eta)$ satisfying $d_t h_t(d_t) = 0$ and $h_t(d_t) + d_t h_t'(d_t) < 0$. Taking the infimum over d_t, we know that $f(g_t) \leq d_t \leq \bar{d} + \eta$. Because $\eta > 0$ is chosen arbitrarily, $f(g_t) \leq \bar{d}$, which in turn implies that there exists some subsequence $f(g_{t_i})$ that has a limit point $\tilde{d}$. Now as long as we can show that $\tilde{d} = d$, the continuity of f can be established. Suppose not, and thus $\tilde{d} < \bar{d}$. However, if so, by the definition of $f(g) = \bar{d}$, it would be the case that $h(\tilde{d}) = 0$, $h'(\tilde{d}) = 0$, or

$$\tilde{d}h(\tilde{d}) = 0, \qquad h(\tilde{d}) + \tilde{d}h'(\tilde{d}) = 0,$$

which is a contradiction to the definition of H. Thus $\tilde{d} = \bar{d}$. Now turn to the case that $f(g) = 0$. So by the definition of f, $0 \cdot h(0) = 0$ and $h(0) + 0 \cdot h'(0) < 0$. Then, by construction, $0 \cdot h_t(0) \to 0$ and $h_t(0) + 0 \cdot h_t'(0) \to h(0) + 0 \cdot h'(0) < 0$. Hence for sufficiently large t, $f(g_t) = f(h_t, h_t') = 0$. The continuity of f in H follows as claimed.

Next, let $\bar{d}_{k,T} = f(l_{k,T}', l_{k,T}'')$, where $(l_{k,T}', l_{k,T}'') \in M_T$. Then by Theorem 1 and the CMT,

$$\bar{d}_{k,T} \to_d \hat{d}_k.$$

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press

But the domain of $\bar{d}_{k,T}$ is not the same as that of $\hat{d}_{k,T}$ (defined on $M_T \cup M_T^0$). We can show that $\Pr(\bar{d}_{k,T} \neq \hat{d}_{k,T})$ is of probability zero in the limit,

$$\begin{aligned}
\lim_{T\to\infty} \Pr(\bar{d}_{k,T} \neq \hat{d}_{k,T}) &\leq \lim_{T\to\infty} (\Pr(\bar{d}_{k,T} \neq \hat{d}_{k,T}, \bar{d}_{k,T} \leq T) + \Pr(\bar{d}_{k,T} > T)) \\
&\leq \lim_{T\to\infty} (\Pr(dl'_{k,T}(d) = dl''_{k,T}(d) + l'_{k,T}(d) = 0 \quad \text{for some } d \geq 0) \\
&\qquad + \Pr(\bar{d}_{k,T} > T)) \\
&= \Pr(dl'_k(d) = dl''_k(d) + l'_k(d) = 0 \quad \text{for some } d \geq 0) \\
&= 0.
\end{aligned}$$

The last and second last equalities are a consequence of applying Lemma A2 and (7), respectively. This establishes the result as claimed. ■

https://doi.org/10.1017/S0266466699151028 Published online by Cambridge University Press