Econometric Theory, **20**, 2004, 116–146. Printed in the United States of America.
DOI: 10.1017/S0266466604201050

EFFICIENT LIKELIHOOD INFERENCE IN NONSTATIONARY UNIVARIATE MODELS

MORTEN ØRREGAARD NIELSEN
Cornell University

Recent literature shows that embedding fractionally integrated time series models with spectral poles at the long-run and/or seasonal frequencies in autoregressive frameworks leads to estimators and test statistics with nonstandard limiting distributions. However, we demonstrate that when embedding such models in a general $I(d)$ framework the resulting estimators and tests regain desirable properties from standard statistical analysis. We show the existence of a local time domain maximum likelihood estimator and its asymptotic normality—and under Gaussianity asymptotic efficiency. The Wald, likelihood ratio, and Lagrange multiplier tests are asymptotically equivalent and chi-squared distributed under local alternatives. With independent and identically distributed Gaussian errors and a scalar parameter, we show that the tests in addition achieve the asymptotic Gaussian power envelope of all invariant unbiased tests; i.e., they are asymptotically uniformly most powerful invariant unbiased against local alternatives. In a Monte Carlo study we document the finite sample superiority of the likelihood ratio test.

1. INTRODUCTION

In this paper we consider likelihood based estimation and testing within a wide class of possibly nonstationary models, including but not limited to the seasonal fractionally integrated autoregressive moving average (ARMA) model. In such models, estimators and test statistics are often found to have nonstandard distributional properties. In contrast, we show that by adapting time domain procedures and embedding the models of interest in a general $I(d)$ framework, instead of the autoregressive alternatives typically considered in the literature, estimators and test statistics regain the standard distributions and optimality properties well known from simpler models.

Several versions of the Wald, likelihood ratio (LR), and score or Lagrange multiplier (LM) testing procedures have appeared in the literature on nonstationary models, e.g., when conducting Dickey and Fuller (1979) type tests for a

I am grateful to Bent Jesper Christensen, Niels Haldrup, Pentti Saikkonen (the co-editor), and two anonymous referees for many useful comments and suggestions that significantly improved this paper. This work was done while the author was at the University of Aarhus, Denmark. Address correspondence to: Morten Ørregaard Nielsen, Department of Economics, Cornell University, 482 Uris Hall, Ithaca, NY 14853; e-mail: mon2@cornell.edu.

  0266-4666/04 $12.00

unit root ($I(1)$ against $I(0)$) or testing stationarity ($I(0)$ against $I(1)$). For a comprehensive recent survey, see Phillips and Xiao (1998). However, these tests have nonstandard limiting distributions that have to be simulated on a case-by-case basis. Some advances have been made recently toward achieving efficient tests. Locally optimal and point optimal tests have been derived for the stationarity hypothesis (e.g., Saikkonen and Luukkonen, 1993a, 1993b) and for the unit root hypothesis (e.g., Elliott, Rothenberg, and Stock, 1996). However, these tests still have nonstandard distributions, and no uniformity results apply.

What is needed is a class of processes that is more general than the unit root $I(1)$ models and admits the testing of smooth hypotheses in the sense that the properties of the process do not differ substantially if the null hypothesis is changed slightly. One such class is that of fractionally integrated processes. Thus, a process is $I(d)$ (fractionally integrated of order d) if its dth difference is $I(0)$; i.e., $y_t \in I(d)$ if

$$(1 - L)^d y_t = e_t \mathbb{I}(t \geq 1), \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and $e_t \in I(0)$. A process is $I(0)$ if it is covariance stationary and its spectrum is bounded and bounded away from zero at any frequency. Testing $H_0: d = 1$ in (1) may be seen as an alternative to unit root testing. We show that in a fractional integration framework much more desirable properties can be obtained compared with autoregressive (and possibly seasonal and fractional) unit root models where test statistics have nonstandard distributions (see, e.g., Phillips, 1987; Hylleberg, Engle, Granger, and Yoo, 1990; Sowell, 1990).

Notable exceptions to the nonstandard tests are Robinson (1994) and Tanaka (1999), extending earlier work by Robinson (1991) and Agiakloglou and Newbold (1994), and it is the Robinson (1994) model that we consider further in this paper. Robinson (1994) derives the LM test statistic (of (16) and (23), which follow) in the frequency domain, claiming that it is more suitable for the analysis, and shows that the LM test statistic is asymptotically chi-squared distributed and locally most powerful under Gaussianity. In a simulation study it is found that when the data generating process (DGP) is of the fractional type the finite sample performance of the new test is better than that of existing tests, the opposite being the case when the DGP is of the autoregressive type. Tanaka (1999) considers the fractional unit root model in (1) and shows the existence of a local time domain maximum likelihood estimator (MLE) and derives the LM and Wald tests. Tanaka (1999) shows that the estimator is asymptotically normal and, under invariance conditions, that the tests are locally most powerful and indeed asymptotically uniformly most powerful against one-sided local alternatives. Simulations show that in finite samples the time domain tests are superior to the frequency domain LM test of Robinson (1994) with respect to both size and power. The estimator is also shown to be quite close to its asymptotic distribution, except in the presence of errors with strong positive serial correlation.

The main contributions of the present paper when compared with the previous work of Robinson (1994) are summarized in the following five points. (i) All the results are obtained in the time domain, which is most frequently employed by practitioners, whereas Robinson (1994) favors a frequency domain approach. The derivation of results and statement of assumptions in the time domain require different methods than in the frequency domain. Another reason to consider the time domain is that in some cases the resulting estimators and tests are more easily applied than their frequency domain counterparts. (ii) It is of interest to examine the estimation of the model by maximum likelihood because the estimator is expected to have good properties. Indeed, it is shown that standard asymptotics and efficiency apply, which is a great advantage in applied work. (iii) Whereas Robinson (1994) only considers the LM test, we also consider the Wald and LR tests and show that standard asymptotics apply to all the test statistics. (iv) For the submodels with a scalar parameter and independent and identically distributed (i.i.d.) Gaussian errors, the LM, LR, and Wald tests are shown to be uniformly most powerful (against local alternatives) among all invariant and unbiased tests. (v) In a simulation study based on the well-known fractional unit root model it is shown that the LR test outperforms the LM and Wald tests with respect to both size and power.

Contrary to the present paper, Tanaka (1999) considers only a special case of the full model in Robinson (1994), namely, the fractional unit root model, and conducts an analysis similar to ours. We consider the full model.

The paper proceeds as follows. In Section 2 we set up the model and discuss important special cases. In Section 3 we consider inference with martingale difference errors and derive the properties of the estimator and tests, whereas in Section 4 we allow serially correlated errors. Section 5 presents the results of our Monte Carlo experiments, and Section 6 concludes. All proofs are collected in the Appendix.

2. THE MODEL

Suppose we observe the real-valued stochastic process $\{y_t, t = 1,2,\ldots,n\}$ generated by the linear model

$$y_t = \beta' x_t + u_t, \tag{2}$$

where $\{x_t\}$ is a $k \times 1$ purely deterministic component and $\{u_t\}$ is an unobserved error component. Two leading cases for the deterministic terms are $x_t = 1$ and $x_t = (1, t)'$, which yield the models $y_t = \beta_0 + u_t$ and $y_t = \beta_0 + \beta_1 t + u_t$, respectively, but other terms such as seasonal dummies are also allowed for; cf. Assumption 2, which follows. The unobserved error process $\{u_t\}$ is assumed to have the generating mechanism

$$\phi(L,\theta)u_t = e_t \mathbb{I}(t \geq 1). \tag{3}$$

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

Here, $\{e_t\}$ is a stationary and invertible process with only weakly dependent errors (i.e., no long memory or nonstationarity) and $\phi(z,\theta)$ is a function of the complex variate z and the $p \times 1$ parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$. The chosen parametrization is such that $\theta = 0$ is the true value, without loss of generality, and this belongs to the interior of Θ.

The model is further required to satisfy the following assumptions.

Assumption 1. The function $\phi(z,\theta)$ is such that (i) $\phi(0,\theta) = 1$ and $\phi(z,\theta) = \phi(z)$ if and only if $\theta = 0$, where $\phi(z) = \phi(z,\theta)|_{\theta=0}$. (ii) $\phi(z,\theta)$ is twice continuously differentiable in θ in an open convex set Θ^* containing Θ and

$$0 < \det(\Psi) < \infty, \tag{4}$$

where $\Psi = \sum_{j=1}^{\infty} \zeta_j \zeta_j'$ and ζ_j is the coefficient on z^j in the expansion of $\zeta(z) = (\partial/\partial\theta) \ln \phi(z,\theta)|_{\theta=0}$ in powers of z. (iii) The function $\lambda(z,\theta) = (\phi(z,\theta)/\phi(z)) \times (\partial/\partial\theta) \ln \phi(z,\theta)$ is continuous in θ at $\theta = 0$ for almost all z such that $|z| = 1$, and, letting $\lambda_j(\theta)$ be the coefficient on z^j in the expansion of $\lambda(z,\theta)$ in powers of z, in a neighborhood N of size $O(n^{-1/2})$ of $\theta = 0$, $\sup_{\theta \in N} \sum_{j=1}^{\infty} \|\lambda_j(\theta)\|^2 < \infty$.

Assumption 2. The $k \times 1$ vector of regressors x_t is nonstochastic and such that $D_n = \sum_{t=1}^{n} \tilde{x}_t \tilde{x}_t'$ is positive definite for n sufficiently large, where $\tilde{x}_t = \phi(L)x_t$.

Assumption 3. The innovation sequence in (3), $\{e_t, t = 0, \pm 1, \pm 2, \ldots\}$, satisfies $E(e_t|\mathcal{F}_{t-1}) = 0$ a.s. and $E(e_t^2|\mathcal{F}_{t-1}) = \sigma^2$ a.s. for all $t \geq 1$, where $\mathcal{F}_t = \sigma(\{e_s, s \leq t\})$ is an increasing sequence of σ-algebrae, and $\{e_t^2, t = 0, \pm 1, \pm 2, \ldots\}$ is uniformly integrable.

Some comments on the assumptions follow. Assumption 1 is a time domain equivalent of the assumptions made by Robinson (1994) on the parametric model, where (i) ensures identifiability of θ and (ii) and (iii) are smoothness conditions on the parametric model. The unit root process nested in an autoregressive model is (3) with $\phi(z,\theta) = 1 - (1 + \theta)z$, but in this case the right-hand-side inequality of (4) is not satisfied. Differentiability to any order is easily verified for all of the examples that follow.

Assumption 2 is a very mild multicollinearity condition on the regressors. It does not even require the smallest eigenvalue of D_n to tend to infinity as $n \to \infty$, which is usually required in linear regression models to get consistent estimates of β.

Finally, Assumption 3 ensures that the innovations are such that $\{e_t, \mathcal{F}_t\}$ and $\{e_t^2 - \sigma^2, \mathcal{F}_t\}$ are both uniformly integrable martingale difference sequences. This is more general than i.i.d. and in practice not much more restrictive than uncorrelatedness. An implication of this assumption is that $n^{-1} \sum_{t=1}^{n} e_t^2 \to \sigma^2$ in probability (e.g., Hall and Heyde, 1980, Theorem 2.22), which we will use later. Assumption 3 can be replaced by any other assumption that gives rise to a weak law of large numbers (LLN) for $\{e_t^2\}$ and a central limit theorem (CLT)

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

in Theorem 3.1, which follows. Thus, we could presumably relax Assumption 3 to accommodate autoregressive conditional heteroskedasticity/generalized autoregressive conditional heteroskedasticity (ARCH/GARCH) type errors (as suggested by an anonymous referee), which are often found in financial data, where our methods are especially applicable due to the large amount of data available (see also Ling and Li, 1997).

A very general model considered by Robinson (1994), and satisfying the preceding assumptions, is

$$\phi(z,\theta) = (1-z)^{d_1+\theta_{i(1)}}(1+z)^{d_2+\theta_{i(2)}}\prod_{j=3}^{h}(1-2\cos\lambda_j z+z^2)^{d_j+\theta_{i(j)}}, \tag{5}$$

where for each j, $\theta_{i(j)} = \theta_l$ for some l, and for each l there is at least one j such that $\theta_{i(j)} = \theta_l$; i.e., there are up to h singularities in the spectral density of u_t and $p \leq h$. That is, we do not require that there is a θ_j for each singularity. For example, the quarterly $I(1)$ hypothesis is given by either one of the functions $\phi(z,\theta) = (1-z^4)^{1+\theta}$, where we use the same θ for each of the $h = 3$ spectral singularities, or $\phi(z,\theta) = (1-z)^{1+\theta_1}(1+z)^{1+\theta_2}(1+z^2)^{1+\theta_3}$, where the integration orders are allowed to be different at different frequencies under the alternative.

The case considered by Tanaka (1999) is the fractional unit root model defined by

$$\phi(z,\theta) = (1-z)^{d+\theta}. \tag{6}$$

In this model $\zeta(z) = \ln(1-z)$ and $\zeta_j = -j^{-1}$ such that $\Psi = \pi^2/6$. The weak dependence, unit root, and $I(2)$ models nested in a fractional integration framework correspond to (6) with $d = 0$, $d = 1$, and $d = 2$, respectively.

Another important special case of the general model (5) is the cyclical $I(d)$ or generalized fractional autoregressive integrated moving average (ARIMA) model of Gray, Zhang, and Woodward (1989), recently advocated by Chung (1996), Bierens (2001), and Gil-Alana (2001). This model is generated by the function

$$\phi(z,\theta) = (1-2\cos\lambda z+z^2)^{d+\theta}, \tag{7}$$

where λ is the cyclic frequency of interest. Then $d = 1$ and $\theta = 0$ corresponds to the cyclic/seasonal unit root at frequency λ.

Finally, suppose the m-vector x_t is $I(d_x)$ and fractionally cointegrated and the cointegrating vector is known a priori from economic theory such that we can treat $u_t = \alpha' x_t$ as an observed time series (when the cointegration vector is unknown and must be estimated, the results in the present paper do not apply; see Nielsen, 2003). Then the hypothesis $H_0\colon \theta = 0$ in (6) with $d = d_x$ corresponds to the null of no fractional cointegration, and with $d = 0$ the hypothesis $H_0\colon \theta = 0$ corresponds to the null of fractional cointegration with $I(0)$ equilibrium errors. A well-known example is the purchasing power parity. Let x_t con-

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

sist of the time t domestic log-price, foreign log-price, and the log exchange rate, respectively, and suppose x_t is fractionally integrated of order d. Then the purchasing power parity predicts that $\alpha = (-1,1,1)'$ should be a cointegrating vector and that the cointegration residuals should be $I(0)$. Imposing $\alpha = (-1,1,1)'$ on the data, the last implication can be tested as in (6) with $d = 0$.

The preceding examples illustrate the generality of our approach. To see why standard asymptotics apply, we briefly discuss the data generating mechanism (see also the discussion by Ling and Li, 2001, pp. 739–741). When $\{u_t\}$ is generated by truncation as in (3), it depends only on the shocks starting at time $t = 1$ and not on shocks starting in the infinite past as would otherwise be the case. Under (3), there are two fundamentally different approaches to allow for nonstationarity that lead to different asymptotic results. Ling and Li (2001) consider the fractional unit root model (6) assuming that $d \in (-\frac{1}{2},\frac{1}{2})$, the stationary region, and allowing unit roots in the autoregressive polynomial $a(z)$. Standard asymptotics is obtained for the fractional difference parameter, but the estimates of the unit roots have nonstandard Dickey–Fuller type distributions. On the other hand, Robinson (1994) and Tanaka (1999) capture the unit root through the fractional difference parameter d and assume that $a(z)$ is stationary. We follow this practice in the present paper. Because no unit root must be estimated in $a(z)$ we avoid the nonsmooth behavior of the model near the unit roots, and this admits standard asymptotics in our setting.

In the subsequent analysis we first consider the case where $\{e_t\}$ is a martingale difference sequence, and then we treat the full model in which $\{e_t\}$ is allowed to follow an ARMA process.

3. INFERENCE WITH MARTINGALE DIFFERENCE ERRORS

The Gaussian log-likelihood function of (2) and (3) is

$$L(\beta,\sigma^2,\theta) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(\phi(L,\theta)(y_t - \beta' x_t))^2 \tag{8}$$

apart from constant terms. The asymptotic results derived in this section impose only Assumption 3 on the error process. Gaussianity is not necessary for most of our results and is used only to choose a likelihood function and to show efficiency.

Because only θ is of interest we concentrate out the nuisance parameter (β',σ^2). This does not influence the results, and in fact $\hat{\theta}$ is asymptotically uncorrelated with $(\hat{\beta}',\hat{\sigma}^2)$ (see the formula for the information matrix (21), which follows). The concentrated likelihood is

$$l(\theta) = L(\beta(\theta),\sigma^2(\theta),\theta) = -\frac{n}{2}\ln(\sigma^2(\theta)), \tag{9}$$

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

where

$$\beta(\theta) = [(\phi(L,\theta)X)'(\phi(L,\theta)X)]^{-1}(\phi(L,\theta)X)'\phi(L,\theta)Y, \tag{10}$$

$$\sigma^2(\theta) = \frac{1}{n}\sum_{t=1}^{n}(\phi(L,\theta)(y_t - \beta(\theta)'x_t))^2, \tag{11}$$

and capital letters denote the appropriate matrices of observations; e.g., X is the $n \times k$ matrix with x_t' as the tth row. Here, $\beta(\theta)$ and $\sigma^2(\theta)$ are functions of θ. They define the estimator $(\hat{\beta}', \hat{\sigma}^2) = (\beta(\hat{\theta}_n)', \sigma^2(\hat{\theta}_n))$ of (β', σ^2). We shall also need $(\tilde{\beta}', \tilde{\sigma}^2) = (\beta(0)', \sigma^2(0))$ which is the estimator of (β', σ^2) under the true value of θ. Note that $\beta(\theta)$ is just the ordinary least squares estimator in a regression of $\phi(L,\theta)y_t$ on $\phi(L,\theta)x_t$ and $\sigma^2(\theta)$ is the usual maximum likelihood variance estimator for the residual process $\phi(L,\theta)(y_t - \beta(\theta)'x_t)$.

Note that the estimate of β need not be consistent in our model. One such case occurs when $x_t = 1$ in the fractional unit root model (6) with $d = 1$. Then $\tilde{\beta} = \beta_0 + e_1$, so $\tilde{\beta}$ is inconsistent, but this has no influence on inference based on $\tilde{u}_t = u_t + (\tilde{\beta} - \beta_0)$ (see Robinson, 1994; also see the Appendix). In fact, what we need in the proofs is the relation

$$E\|(\tilde{\beta} - \beta)'D_n^{1/2}\| = O(1), \tag{12}$$

which follows under Assumption 2 by definition of $\tilde{\beta}$.

3.1. Estimation

In this section we show the existence of a local MLE and derive the limiting distribution theory following the approach of Sargan and Bhargava (1983) and Tanaka (1999). In particular, we consider the conditional sum of squared residuals objective function (9).

In the following we find it convenient to consider maximizing

$$g(\theta) = l(\theta) - l(0) = -\frac{n}{2}\ln\left[1 - \frac{1}{n}\frac{\sum_{t=1}^{n}(\phi(L)\tilde{u}_t)^2 - \sum_{t=1}^{n}(\phi(L,\theta)\hat{u}_t)^2}{\frac{1}{n}\sum_{t=1}^{n}(\phi(L)\tilde{u}_t)^2}\right], \tag{13}$$

where $\tilde{u}_t = y_t - \tilde{\beta}'x_t$ and $\hat{u}_t = y_t - \hat{\beta}'x_t$. Assume first that we are in a neighborhood of the true value, i.e., that there exists a δ such that $\theta = \delta/\sqrt{n}$ (the existence of δ will be proved shortly). Then we can show the following results.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

THEOREM 3.1. *Let Assumptions 1–3 be satisfied and let $g(\theta)$ be given by (13). Then, under $\theta = \delta/\sqrt{n}$,*

$$(i)\ g(\theta) \to_d W(\delta) = \frac{\delta'}{2}(2\Psi^{1/2}Z - \Psi\delta),$$

$$(ii)\ \frac{\partial g(\theta)}{\partial \delta} \to_d \frac{\partial W(\delta)}{\partial \delta} = \Psi^{1/2}Z - \Psi\delta,$$

$$(iii)\ \frac{\partial^2 g(\theta)}{\partial \delta \partial \delta'} \to_p -\Psi,$$

where Z is a p-dimensional standard normal random vector.

Next, we prove the existence of a local MLE $\hat{\theta}_n$ of $\theta_0 = 0$ such that $\sqrt{n}\hat{\theta}_n = \hat{\delta} = O_p(1)$ following Sargan and Bhargava (1983) and Tanaka (1999). Let ι be a $p \times 1$ direction vector, i.e., satisfying $\|\iota\| = 1$, where $\|\cdot\|$ is the Euclidean norm, and let $\delta = \|\delta\|\iota$. Generalizing the scalar approach by Sargan and Bhargava (1983) and Tanaka (1999), it suffices to show that

$$P\left(\iota' \frac{\partial g(\delta/\sqrt{n})}{\partial \delta} \geq 0\right) \leq \varepsilon \tag{14}$$

for any direction vector ι, $\varepsilon > 0$, and $n \geq n_0$ (n_0 fixed) and for some $\|\delta\| > 0$. Note that $\iota' \partial g(\delta/\sqrt{n})/\partial\delta$ is the directional derivative of g at $\delta/\sqrt{n}$, i.e., the rate of change of g at $\delta/\sqrt{n}$ in the direction ι.

Thus, for all direction vectors ι, moving some distance $\|\delta\|$ in the direction ι from the true value, the directional derivative of g in the same direction ι should be negative for sufficiently large n. In the one-dimensional case $\iota = \pm 1$ and (14) reduces to the corresponding conditions of Sargan and Bhargava (1983) and Tanaka (1999). It follows from Theorem 3.1 that

$$\begin{aligned} P\left(\iota' \frac{\partial g(\delta/\sqrt{n})}{\partial \delta} \geq 0\right) &\to P\left(\iota' \frac{\partial W(\delta)}{\partial \delta} \geq 0\right) \\ &= P\left(\iota' \frac{\partial W(\delta)}{\partial \delta} - E\iota' \frac{\partial W(\delta)}{\partial \delta} \geq -E\iota' \frac{\partial W(\delta)}{\partial \delta}\right) \\ &\leq \frac{\mathrm{Var}\left(\iota' \dfrac{\partial W(\delta)}{\partial \delta}\right)}{\left(E\iota' \dfrac{\partial W(\delta)}{\partial \delta}\right)^2} \\ &= \frac{1}{\iota'\Psi\iota\|\delta\|^2}, \end{aligned}$$

which can be made arbitrarily small by selecting $\|\delta\|$ large. Thus, (14) holds by appropriate choices of $\|\delta\|$ and n_0, and the existence of the local MLE $\hat{\theta}_n$ is ensured.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

THEOREM 3.2. *Under Assumptions 1–3, there exists a local maximizer* $\hat{\theta}_n$ *of the concentrated likelihood (9) that satisfies, as* $n \to \infty$,

$$\sqrt{n}\hat{\theta}_n \to_d N(0, \Psi^{-1}), \tag{15}$$

and under the additional assumption of Gaussianity of $\{e_t\}$, $\hat{\theta}_n$ *is asymptotically efficient in the sense that its asymptotic variance attains the Cramér–Rao lower bound.*

This asymptotic normality result stands in sharp contrast, e.g., to the nonstandard Dickey–Fuller distribution. In that case, $n^{-1}\partial l(\theta)/\partial\theta|_{\theta=0} \Rightarrow \frac{1}{2}(W(1)-1)^2$, $n^{-2}\partial^2 l(\theta)/\partial\theta\partial\theta' \Rightarrow \int_0^1 W^2(t)\,dt$, and thus $n\hat{\theta} \Rightarrow \frac{1}{2}(W(1)-1)^2/\int_0^1 W^2(t)\,dt$, where $W(t)$ is a standard Brownian motion and $\Rightarrow$ is weak convergence (see, e.g., Phillips, 1987; Phillips and Xiao, 1998). Furthermore, if a constant term is included in the Dickey–Fuller model the distribution changes. This is not the case in our model, where the limiting distribution is independent of the nuisance parameter (β', σ^2).

The additional assumption of Gaussianity allows a strengthening of the results. Thus, $\hat{\theta}_n$ is asymptotically the best estimator in the class of all $\sqrt{n}$-consistent and asymptotically normal estimators. This result also is in contrast with those usually found in the theory of nonstationary time series.

The simple asymptotic distribution in Theorem 3.2 makes it easy to construct p-dimensional confidence ellipsoids for θ or conduct Wald-type tests of hypotheses on θ. This is examined in detail in the next section.

3.2. Hypothesis Testing

Suppose we wish to test the hypothesis

$$H_0 : \theta = \theta_0 = 0, \tag{16}$$

where θ_0 is set to zero because otherwise we would get trivial asymptotic distributions under the null. Robinson (1994) considers the LM test in a frequency domain framework. We now consider all the classical likelihood-based (Wald, LR, LM) tests (see Engle, 1984) in the time domain.

From Theorem 3.2, the Wald test statistic is

$$W = n\hat{\theta}_n'\Psi\hat{\theta}_n. \tag{17}$$

We denote by a tilde an estimator under the null hypothesis. The (quasi) LR test statistic is given by

$$LR = 2(L(\hat{\beta}, \hat{\sigma}^2, \hat{\theta}_n) - L(\tilde{\beta}, \tilde{\sigma}^2, 0)) = n \ln\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right) \tag{18}$$

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

(see equation (9)). Finally, to derive the LM test statistic

$$LM = \frac{\partial L(\eta)}{\partial \eta'}\left[E_0\left(\frac{\partial L(\eta)}{\partial \eta}\,\frac{\partial L(\eta)}{\partial \eta'}\right)\right]^{-1}\frac{\partial L(\eta)}{\partial \eta}\bigg|_{\beta=\tilde{\beta},\sigma^2=\tilde{\sigma}^2,\theta=0}, \tag{19}$$

where $\eta = (\beta', \sigma^2, \theta')'$, we note that

$$\frac{\partial L(\beta,\sigma^2,\theta)}{\partial \theta}\bigg|_{\beta=\tilde{\beta},\sigma^2=\tilde{\sigma}^2,\theta=0} = \tilde{\sigma}^{-2}\sum_{t=1}^{n}\sum_{j=1}^{t-1}\zeta_j\,\tilde{e}_{t-j}\,\tilde{e}_t = n\tilde{A}_n, \tag{20}$$

whereas the other two partial derivatives vanish. Here, $\tilde{A}_n = \sum_{j=1}^{n-1}\zeta_j\,\tilde{\rho}(j)$ and $\tilde{\rho}(j)$ is the jth sample autocorrelation of $\tilde{e}_t = \phi(L)(y_t - \tilde{\beta}' x_t)$.

The diagonal block of the Fisher information matrix corresponding to θ is

$$\begin{aligned}\frac{1}{\sigma^4}\sum_{t=1}^{n}\sum_{s=1}^{n}\sum_{j=1}^{t-1}\sum_{i=1}^{s-1}\zeta_j\zeta_i' E_0(e_{t-j}\,e_t\,e_s\,e_{s-i}) &= \frac{1}{\sigma^4}\sum_{t=1}^{n}\sum_{j=1}^{t-1}\zeta_j\zeta_j' E(e_{t-j}^2 E(e_t^2|\mathcal{F}_{t-1})) \\ &= n\sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)\zeta_j\zeta_j',\end{aligned}$$

so the Fisher information matrix in (19) evaluated at $\beta = \tilde{\beta}, \sigma^2 = \tilde{\sigma}^2, \theta = 0$ is n times

$$\begin{bmatrix}\sigma^{-2}D_n & 0 & 0\\ 0 & \frac{1}{2}\sigma^{-4} & 0\\ 0 & 0 & \Psi_n\end{bmatrix}, \tag{21}$$

which is invertible for n sufficiently large by (4) and Assumption 2. The diagonal blocks corresponding to β and σ^2 follow using that $\{e_t, \mathcal{F}_t\}$ and $\{e_t^2 - \sigma^2, \mathcal{F}_t\}$ are martingale differences, respectively. In Tanaka (1999), $\zeta_j = j^{-1}$ and $\Psi = \pi^2/6$. We allow for more general weights to the autocorrelations in $\tilde{A}_n$, corresponding to the more flexible model represented by the function $\phi(z,\theta)$. The expression $\Psi_n = \sum_{j=1}^{n-1}(1 - (j/n))\zeta_j\zeta_j'$ is a truncated version of Ψ, which is asymptotically equivalent to Ψ. Thus, the LM test statistic is

$$LM = n\tilde{A}_n'\Psi^{-1}\tilde{A}_n. \tag{22}$$

In the fractional unit root model (6) where $\zeta_j = j^{-1}$ we have $\Psi_{100} = 1.5831$, $\Psi_{500} = 1.6294$, and $\Psi = \Psi_\infty = \pi^2/6 = 1.6449$.

We derive the distribution of the test statistics under the more general assumption of local (Pitman) alternatives given by the sequence

$$H_1 : \theta = \theta_{1n} = \delta/\sqrt{n} \tag{23}$$

with δ a fixed $p \times 1$ vector.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

THEOREM 3.3. *Let Assumptions 1–3 be satisfied and let T denote the* W, LR, *or* LM *test statistics given by (17), (18), and (22). Then, under (23), it holds that*

$$T \to_d \chi_p^2(\delta' \Psi \delta)$$

as $n \to \infty$. *The three tests are consistent and asymptotically equivalent; i.e., if* T_1 *and* T_2 *are any two of the statistics then* $T_1 - T_2 \to 0$ *in probability. Under the additional assumption of Gaussianity they are locally most powerful.*

Usually in nonstandard tests such as the Dickey–Fuller test, the three test statistics are not equivalent. From the proof we note that the equivalence of the tests depends crucially on the information matrix equality, which holds asymptotically in our model but does not hold when the unit root is nested in an autoregressive alternative.

Thus, we find that unusually simple asymptotic tests can be performed in this model using the chi-squared distribution. Also, we can easily calculate the asymptotic local power of the three test statistics, which we state as a corollary.

COROLLARY 3.1. *Under the conditions of Theorem 3.3 it holds that, under* $\theta = \delta/\sqrt{n}$,

$$P(T > \chi_{p,1-\alpha}^2) \to 1 - F_{\delta' \Psi \delta}(\chi_{p,1-\alpha}^2) \tag{24}$$

as $n \to \infty$, *where* $\chi_{p,1-\alpha}^2$ *is the* $100(1-\alpha)\%$ *point of the* χ_p^2 *distribution and* $F_{\delta' \Psi \delta}$ *is the distribution function of the* $\chi_p^2(\delta' \Psi \delta)$ *distribution.*

Using Corollary 3.1 we can compare the finite sample performance of the tests with the approximation offered by asymptotic theory, and we shall discuss this in Section 5.

Next, we show that even stronger results can be obtained in a subclass of models.

3.3. Uniformly Most Powerful Tests

While the general theory discussed previously applies for multidimensional θ, even stronger results are obtained in the special case of scalar θ, e.g., (6) or (7), which we now consider briefly. Following the reasoning in Elliott et al. (1996) and Tanaka (1999), we derive the power envelope for the two-sided testing problems under invariance and unbiasedness conditions and show that this two-sided power envelope is equal to (24), i.e., that this power is achieved by our tests. The unbiasedness condition is new because Elliott et al. (1996) and Tanaka (1999) only consider one-sided tests and thus do not need unbiasedness.

In particular, we assume that the errors are Gaussian and that the model in (3) is characterized by a scalar parameter θ. This rules out the general model

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

in (5) but still applies to most of the models in Section 2. The testing problem is invariant to any transformation of the type $y \to ay + Xb$ ($a > 0$ and $b \in \mathbb{R}^k$), or in the parameter space,

$$(\theta, \beta, \sigma^2) \to (\theta, b + a\beta, a^2\sigma^2). \tag{25}$$

Thus, we shall restrict attention to the family of tests that are invariant to the group of transformations in (25) (see Lehmann, 1986, Chap. 6).

Assume that the DGP is given by (2) and (3), with true parameter value $\theta_{0n} = c/\sqrt{n}$ for some fixed c. Now consider testing the hypothesis $H_0: \theta = 0$ against the sequence of local alternatives $H_1: \theta_{1n} = \delta/\sqrt{n}$ for some fixed δ. This is a test of a simple null vs. a simple alternative with nuisance parameter (β', σ^2). Then we can apply invariance arguments to (β', σ^2) and the Neyman–Pearson lemma tells us (e.g., Lehmann, 1986, p. 338) that the test that rejects the null when

$$M_n = n \frac{\sum_{t=1}^{n} \tilde{e}_{tn}^2 - \sum_{t=1}^{n} \hat{e}_{tn}^2}{\sum_{t=1}^{n} \tilde{e}_{tn}^2} \tag{26}$$

becomes large is most powerful invariant (MPI) with respect to the group of transformations (25). As in the previous section, $\tilde{e}_{tn}$ and $\hat{e}_{tn}$ are residuals under H_0 and H_1, respectively. The next theorem derives the limiting distribution of M_n under local alternatives.

THEOREM 3.4. *Let M_n denote the MPI test statistic (26), with $\theta_{0n} = c/\sqrt{n}$ (c a fixed scalar) instead of $\theta_0 = 0$. Let Assumptions 1 and 2 be satisfied and suppose the error process is i.i.d. Gaussian. Then, under the sequence of local alternatives $\theta_{1n} = \delta/\sqrt{n}$ (δ a fixed scalar), it holds that*

$$M_n \to_d M(c, \delta) = 2\delta\sqrt{\Psi}Z + \delta(2c - \delta)\Psi$$

as $n \to \infty$, where Z is a standard normal variable.

Thus, invariance arguments have reduced the testing problem to the consideration of the statistic M_n, and the power envelope of all invariant tests is the power of $M(\delta, \delta)$. Obviously, the results in Tanaka (1999) apply with little change to the corresponding one-sided testing problem in our setup and this power envelope is achieved by one-sided versions of our tests. However, because we consider mainly the two-sided testing problem, we cannot hope to achieve the same power envelope, and thus the following results differ from those in Tanaka (1999), where only one-sided hypotheses are considered.

To find a test statistic that applies against two-sided alternatives we invoke the principle of unbiasedness (see Lehmann, 1986, Ch. 4) to construct an MPI unbiased test. Unbiasedness requires that the power of the test never falls be-

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

low the nominal significance level for any point in the alternative. Because for varying c the family of distributions $M(c,\delta)$ is normal, it satisfies the requirement that it be strictly totally positive of order three (STP_3; see Lehmann, 1986, p. 119), and hence the power envelope of all invariant and unbiased tests of $H_0\colon\theta = 0$ against $H_1\colon\theta_{1n} = \delta/\sqrt{n}$ is given by $\Pi(\delta) = 1 - P(C_{1,\alpha}(\delta) < M(\delta,\delta) < C_{2,\alpha}(\delta))$ (Lehmann, 1986, p. 303), where the constants are determined by

$$P(C_{1,\alpha}(\delta) < M(0,\delta) < C_{2,\alpha}(\delta)) = 1 - \alpha, \tag{27}$$

$$\left.\frac{\partial P(C_{1,\alpha}(\delta) < M(c,\delta) < C_{2,\alpha}(\delta))}{\partial c}\right|_{c=0} = 0. \tag{28}$$

A test whose asymptotic power attains the power envelope for all points δ is asymptotically uniformly most powerful invariant unbiased. The following theorem shows that the power envelope of all invariant and unbiased tests is given by (24), i.e., that this power is achieved by our tests.

THEOREM 3.5. *Let Assumptions 1 and 2 be satisfied and suppose the error process is i.i.d. Gaussian. Then the asymptotic Gaussian power envelope of all invariant (with respect to (25)) and unbiased tests of* $H_0\colon\theta = 0$ *against* $H_1\colon\theta_{1n} = \delta/\sqrt{n}$ *(*δ *a fixed scalar) is given by (24). Thus, the* W, LR, *and* LM *tests are uniformly most powerful (against local alternatives) among all invariant and unbiased tests.*

This result is in stark contrast to the results of Saikkonen and Luukkonen (1993a, 1993b) and Elliott et al. (1996), among others, whose tests are only point optimal invariant, i.e., tests that have maximal power against a single prespecified (local) point in the alternative. Our criterion is against all possible (local) alternatives.

Furthermore, Theorem 3.5 also applies to the test statistic in Robinson (1994) and thus generalizes his result, too, because he only shows that his test is locally most powerful.

4. INFERENCE WITH SERIALLY CORRELATED ERRORS

Now we extend the basic model to allow for weakly dependent (ARMA) errors. In particular, we work with the following assumption.

Assumption 4. $\{e_t\}$ is generated by an ARMA model of the form

$$a(L)e_t = b(L)\varepsilon_t, \tag{29}$$

where $\{\varepsilon_t\}$ satisfies Assumption 3. Here $a(z)$ and $b(z)$ are finite polynomials without common roots and all roots strictly outside the unit circle. The coefficients in the autoregressive and moving average polynomials are collected in the $q \times 1$ parameter vector ψ.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

This assumption follows Tanaka (1999); Tanaka in addition assumes that $\{\varepsilon_t\}$ is i.i.d. Thus, we offer more generality in this respect too, because of our martingale difference assumption on $\{\varepsilon_t\}$.

Collect the parameters of the dynamic part of the model in the vector $\gamma = (\theta', \psi')'$ with true value $\gamma_0 = (0', \psi_0')'$ and let $c(z,\psi) = a(z)b^{-1}(z)$. Analogously to $\zeta(z,\theta)$, define $\xi(z,\gamma) = \partial \ln(\phi(z,\theta)c(z,\psi))/\partial\gamma$ and $\xi(z) = \partial \ln(\phi(z,\theta)c(z,\psi))/\partial\gamma|_{\gamma=\gamma_0} = \sum_{j=1}^{\infty} \xi_j z^j$. Note that $\xi_j = (\zeta_j', c_j')'$ with ζ_j defined as before and c_j defined as the coefficient on z^j in the expansion of $\partial \ln c(z,\psi)\partial\psi|_{\psi=\psi_0}$ in powers of z. As in Assumption 1(ii) we define

$$\Xi = \sum_{j=1}^{\infty} \xi_j \xi_j' = \begin{bmatrix} \Psi & \kappa' \\ \kappa & \Phi \end{bmatrix} \tag{30}$$

with $\kappa = \sum_{j=1}^{\infty} c_j \zeta_j'$ and $\Phi = \sum_{j=1}^{\infty} c_j c_j'$.

It is easily shown that Φ is the Fisher information for ψ under Assumption 4; e.g., if $\{e_t\}$ is an AR(1) process with coefficient a then $c_j = -a^{j-1}$ and $\Phi = (1-a^2)^{-1}$. Finally, corresponding to (4), we assume that

$$0 < \det(\Psi - \kappa'\Phi^{-1}\kappa) < \infty, \tag{31}$$

which in particular implies that Ξ is nonsingular.

The log-likelihood function in the case of serially correlated errors is, except for constants,

$$L(\beta, \sigma^2, \gamma) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(\phi(L,\theta)c(L,\psi)(y_t - \beta' x_t))^2, \tag{32}$$

to be compared with (8). The concentrated likelihood function for $\gamma = (\theta', \psi')'$ becomes

$$l(\gamma) = -\frac{n}{2}\ln(\sigma^2(\gamma)) \tag{33}$$

except for constants, where

$$\begin{aligned}\beta(\gamma) = {} & [(\phi(L,\theta)c(L,\psi)X)'(\phi(L,\theta)c(L,\psi)X)]^{-1} \\ & \times (\phi(L,\theta)c(L,\psi)X)'\phi(L,\theta)c(L,\psi)Y,\end{aligned} \tag{34}$$

$$\sigma^2(\gamma) = \frac{1}{n}\sum_{t=1}^{n}(\phi(L,\theta)c(L,\psi)(y_t - \beta(\gamma)' x_t))^2, \tag{35}$$

and $(\hat\beta', \hat\sigma^2)$ and $(\tilde\beta', \tilde\sigma^2)$ are now defined in terms of the functions (34) and (35). Corresponding to (13) we consider the function

$$\begin{aligned} g(\gamma) &= l(\gamma) - l(\gamma_0) \\ &= -\frac{n}{2}\ln\left[1 - \frac{1}{n}\frac{\sum_{t=1}^{n}(\phi(L)c(L,\psi_0)\tilde u_t)^2 - \sum_{t=1}^{n}(\phi(L,\theta)c(L,\psi)\hat u_t)^2}{\frac{1}{n}\sum_{t=1}^{n}(\phi(L)c(L,\psi_0)\tilde u_t)^2}\right]. \end{aligned}$$

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

4.1. Estimation

The analysis of the model with serially correlated errors proceeds in the same way as with martingale difference errors as discussed previously. Thus, we are able to show the existence of a local MLE $\hat{\gamma}_n = (\hat{\theta}_n', \hat{\psi}_n')'$ satisfying $\sqrt{n}\hat{\gamma}_n = O_p(1)$ and to prove joint asymptotic normality of $\hat{\theta}_n$ and $\hat{\psi}_n$. Under Gaussianity we achieve efficiency as before.

THEOREM 4.1. *Under Assumptions 1, 2, and 4 and (31) there exists a local maximizer $\hat{\gamma}_n = (\hat{\theta}_n', \hat{\psi}_n')'$ of the concentrated likelihood (33) that satisfies, as $n \to \infty$,*

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \to_d N(0, \Xi^{-1}). \tag{36}$$

Under the additional assumptions of Gaussianity of $\{\varepsilon_t\}$ and correct (minimal) specification (all elements of ψ_0 are nonzero), $\hat{\gamma}_n$ is asymptotically efficient in the sense that its asymptotic variance attains the Cramér–Rao lower bound.

Based on this theorem it is possible to create joint $(p + q)$-dimensional confidence ellipsoids for θ and ψ that take into account the asymptotic correlation between the estimates represented by the matrix κ. This is important for inference, not only on θ but also on ψ. Usually, in applied work one would determine the appropriate filtration of data (i.e., the function $\phi(z, \theta)$) by Dickey–Fuller tests or similar methods and then treat the filtered data as if it were observed, i.e., as if the correct filter were known a priori. The resulting inference on ψ is incorrect, because the correlation between θ and ψ is ignored. When applying Theorem 4.1, this pretesting problem is avoided because θ and ψ are estimated jointly.

When inference on θ is of interest, the asymptotic marginal distribution of $\hat{\theta}_n$ can be immediately derived from the theorem.

COROLLARY 4.1. *Under the conditions of Theorem 4.1,*

$$\sqrt{n}\hat{\theta}_n \to_d N(0, (\Psi - \kappa'\Phi^{-1}\kappa)^{-1}) \tag{37}$$

as $n \to \infty$.

In parallel with Corollary 4.1, $\mathrm{var}(\sqrt{n}(\hat{\psi}_n - \psi_0)) \to (\Phi - \kappa\Psi^{-1}\kappa')^{-1}$ (by the partitioned matrix inverse formula), and in the special case where ϕ is not present this reduces to Φ^{-1}, which is the Fisher information on ψ. Thus, the well-known asymptotic efficiency of the MLE in pure ARMA models comes out as a special case of our results. More important, Theorem 4.1 with ϕ present demonstrates the joint efficiency in the generalized model.

To illustrate the loss of efficiency in estimation of θ stemming from serially correlated errors, consider again the fractional unit root model. Suppose we know that the errors are not serially correlated but simply are martingale differences. Then the asymptotic variance of $\sqrt{n}\hat{\theta}_n$ is $6/\pi^2$ by Theorem 3.2. If

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

instead it is known that the errors exhibit serial correlation of the AR(1) or MA(1) type with coefficient a, then the asymptotic variance of $\sqrt{n}\hat{\theta}_n$ is the inverse of $(\pi^2/6) - ((1-a^2)/a^2)(\ln(1-a))^2$ by Corollary 4.1.

Figure 1 shows the relative efficiency of these two estimates as a function of the serial correlation parameter, a. This is calculated as

$$1 - \frac{6}{\pi^2}\frac{1-a^2}{a^2}(\ln(1-a))^2, \tag{38}$$

which has a minimum at $a = 0.684$. This suggests that moderate levels of a best replicate the behavior of the (weighted) autocorrelations of a fractionally integrated process. The point $a = 0$ shows that the relative efficiency allowing for serial correlation when it is not present is 0.392, as noted by Tanaka (1999).

4.2. Hypothesis Testing

We now consider the testing problems (16) and (23) in the presence of serially correlated errors, where again only θ is of interest. The Wald, LR, and LM test statistics are

$$W = n\hat{\theta}_n'(\Psi - \hat{\kappa}'\hat{\Phi}^{-1}\hat{\kappa})\hat{\theta}_n, \tag{39}$$

$$LR = n\ln\left(\frac{\sigma^2(0,\tilde{\psi}_n)}{\sigma^2(\hat{\theta}_n,\hat{\psi}_n)}\right), \tag{40}$$

$$LM = n\tilde{A}_n'(\Psi - \tilde{\kappa}'\tilde{\Phi}^{-1}\tilde{\kappa})^{-1}\tilde{A}_n, \tag{41}$$

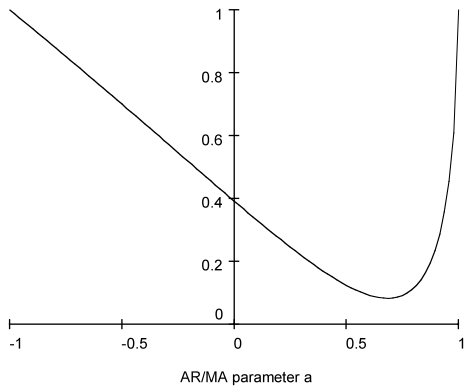

FIGURE 1. Relative efficiency of $\hat{\theta}_n$ in the presence of first-order autoregressive or moving average errors.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

where $\hat{\kappa}$ and $\hat{\Phi}$ are evaluated at $\hat{\psi}_n$ and $\tilde{\kappa}$ and $\tilde{\Phi}$ are evaluated at $\tilde{\psi}_n$, the estimate of ψ under the null, and $\tilde{A}_n = \sum_{j=1}^{n-1} \zeta_j \tilde{\rho}(j)$ is defined in terms of the jth sample autocorrelation of $\tilde{\varepsilon}_t = \phi(L)c(L, \tilde{\psi}_n)(y_t - \tilde{\beta}' x_t)$.

It is obvious from the expressions for the test statistics and $\tilde{A}_n$ that the LM test is not necessarily the simplest to apply in practice. The implementation of the Wald and LR test statistics is straightforward when we can estimate the model under both the null and alternative and should not be a problem, given the methods available in the previous sections. In particular, the LR test is attractive because there is no need to calculate Ψ, κ, and Φ.

Similar to the calculation of the infinite-order moving average coefficients in standard ARMA models, the calculation of κ and Φ can be quite cumbersome when the model in Assumption 4 is more complex than just an AR(1) or MA(1) model (see also the discussion in Tanaka, 1999). To overcome this issue, one could employ the numerical approximations

$$\hat{W} = n\hat{\theta}_n' \left(\sum_{t=1}^{n} \frac{\partial \hat{\varepsilon}_t}{\partial \theta} \frac{\partial \hat{\varepsilon}_t}{\partial \theta'} \Big/ \sum_{t=1}^{n} \hat{\varepsilon}_t^2 \right) \hat{\theta}_n \Bigg|_{H_1},$$

$$\widehat{LM} = n \sum_{t=1}^{n} \tilde{\varepsilon}_t \frac{\partial \tilde{\varepsilon}_t}{\partial \theta'} \left(\sum_{t=1}^{n} \frac{\partial \tilde{\varepsilon}_t}{\partial \theta} \frac{\partial \tilde{\varepsilon}_t}{\partial \theta'} \sum_{t=1}^{n} \tilde{\varepsilon}_t^2 \right)^{-1} \sum_{t=1}^{n} \tilde{\varepsilon}_t \frac{\partial \tilde{\varepsilon}_t}{\partial \theta} \Bigg|_{H_0},$$

which of course have the same asymptotic properties as W and LM. However, because Ψ, κ, and Φ, and thus W and LM, can be calculated for any given parameter value (say, $\bar{\gamma}$) by numerically expanding $\partial \ln \phi(z, \theta) c(z, \psi)/\partial \gamma$ at $\gamma = \bar{\gamma}$ in powers of z using a computer, we do not consider $\hat{W}$ and $\widehat{LM}$ further.

The asymptotic distribution of the tests under local alternatives and with serial correlation is given by the following theorem.

THEOREM 4.2. *Let Assumptions 1, 2, and 4 and (31) be satisfied and let T denote the* W, LR, *or* LM *test statistics (39), (40), and (41). Then, under (23), it holds that*

$$T \to_d \chi_p^2(\delta'(\Psi - \kappa' \Phi^{-1} \kappa)\delta)$$

as $n \to \infty$. *The three tests are consistent and asymptotically equivalent, and under the additional assumption of Gaussianity they are locally most powerful.*

This theorem shows that the tests are still locally most powerful, even in the presence of serially correlated errors. Setting $\kappa = \Phi = 0$, i.e., when no serial correlation is present and ψ is not estimated, generates Theorem 3.3 as a special case. As with Corollary 3.1 in the case without serial correlation, we can easily calculate the asymptotic local power, giving us a benchmark against which to compare the power of the tests in finite samples.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

COROLLARY 4.2. *Under the conditions of Theorem 4.2 it holds that, under* $\theta = \delta/\sqrt{n}$,

$$P(T > \chi^2_{p,1-\alpha}) \to 1 - F_{\delta'(\Psi-\kappa'\Phi^{-1}\kappa)\delta}(\chi^2_{p,1-\alpha}) \tag{42}$$

as $n \to \infty$, *where* $\chi^2_{p,1-\alpha}$ *is the* $100(1-\alpha)\%$ *point of the* χ^2_p *distribution and* $F_{\delta'(\Psi-\kappa'\Phi^{-1}\kappa)\delta}$ *is the distribution function of the* $\chi^2_p(\delta'(\Psi - \kappa'\Phi^{-1}\kappa)\delta)$ *distribution.*

Using Corollary 4.2, Figure 2 shows the local power functions against positive alternatives for the fractional unit root model with different specifications of AR(1) errors. Because δ only enters (42) through δ^2, the power functions are symmetric. The starred line is the local power function when the errors are a martingale difference sequence and this is known (i.e., using Corollary 3.1). The dotted, dashed, and solid lines correspond to AR(1) specifications of the errors with coefficient $a = -0.5$, $a = 0$, and $a = 0.5$, respectively. In the case $a = 0$, the errors are a martingale difference sequence, but an AR(1) error process is estimated.

The local power of the tests in the model with $a = 0.5$ is much lower than for the other specifications. On the other hand, the power loss in the model with $a = -0.5$ is small. This is in accordance with the results in Section 4.1; cf. (38) and Figure 1.

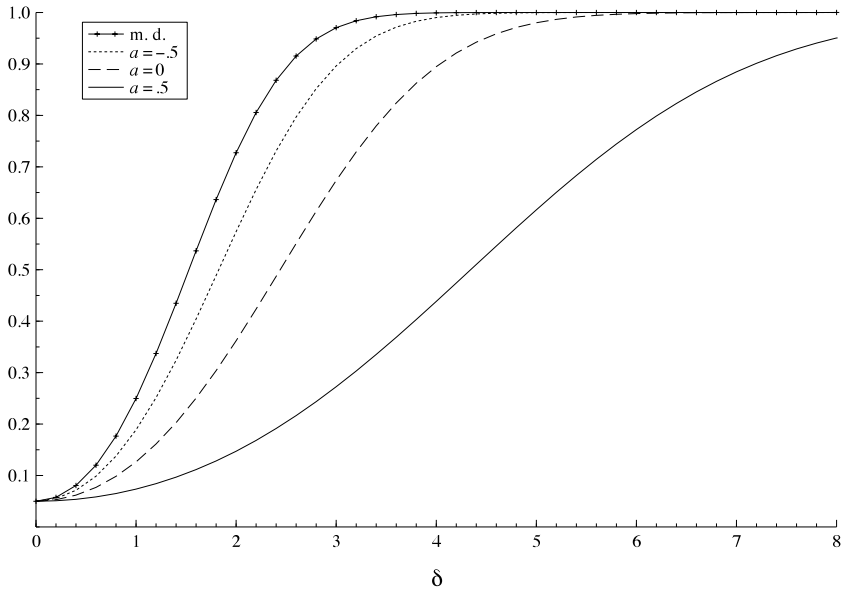

FIGURE 2. Asymptotic local power functions with martingale difference and first-order autoregressive or moving average errors.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

5. FINITE SAMPLE PERFORMANCE

In this section, we compare the asymptotic local power functions derived in the previous sections to the finite sample rejection frequencies by means of Monte Carlo experiments.

The model we use for the simulation study is the well-known fractional unit root model with an AR(1) error:

$$(1-L)^{1+\theta}y_t = e_t \mathbb{I}(t \geq 1), \tag{43}$$

$$(1-aL)e_t = \varepsilon_t, \tag{44}$$

where $\{\varepsilon_t\}$ is i.i.d. standard normal. This model is also studied in simulations by Robinson (1994) and Tanaka (1999). In addition to this fractional DGP, Robinson (1994) also considers an autoregressive DGP and finds that his test is dominated by Dickey–Fuller type tests in the latter case.

We concentrate on comparing the finite sample performance of the three test statistics (Wald, LR, and LM). Tanaka (1999) documents that the time domain LM test outperforms Robinson's (1994) frequency domain LM test, so we do not consider the frequency domain test here. The properties of the estimator $\hat{\theta}_n$ in this model are examined by Tanaka (1999), who finds that in the case without serial correlation the behavior of the local MLE is very close to the asymptotic distribution. However, with serially correlated errors the performance of the local MLE degrades, and especially in the case of strong positive serial correlation the performance is poor. This is expected based on (38) and Figure 1.

Throughout, we fix the nominal level (type I error) at 0.05 and the number of replications at 5,000. We consider the sample sizes $n = 100$ and $n = 500$. The former is typical for macroeconomic time series and the latter (or even larger) for financial time series. For each experiment, 5,000 samples of size $n = 500$ were generated using the rann, diffpow, and armagen routines in Ox version 3.00 including the Arfima package version 1.01 (see Doornik, 2001; Doornik and Ooms, 2001). For the smaller sample size, $n = 100$, we used the first 100 out of the 500 observations from each sample.

Figures 3–6 present the simulated finite sample power functions of the test statistics for different specifications of the error term in (44) (the tables containing the numerical values used to construct the figures can be obtained from the author upon request). For each value of θ, the asymptotic local power has been calculated by setting $\delta = \theta\sqrt{n}$ in Corollaries 3.1 and 4.2 and is reported under the heading Limit. In all the figures, the left-hand-side figures (a) and (c) present the simulated power functions of the tests calculated as in Sections 3.2 and 4.2, whereas the simulated power functions in the right-hand-side figures (b) and (d) are calculated using size corrected critical values.

First, consider the case of martingale difference errors shown in Figure 3, i.e., $\{e_t\} = \{\varepsilon_t\}$. In this case, all the finite sample rejection frequencies are very

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

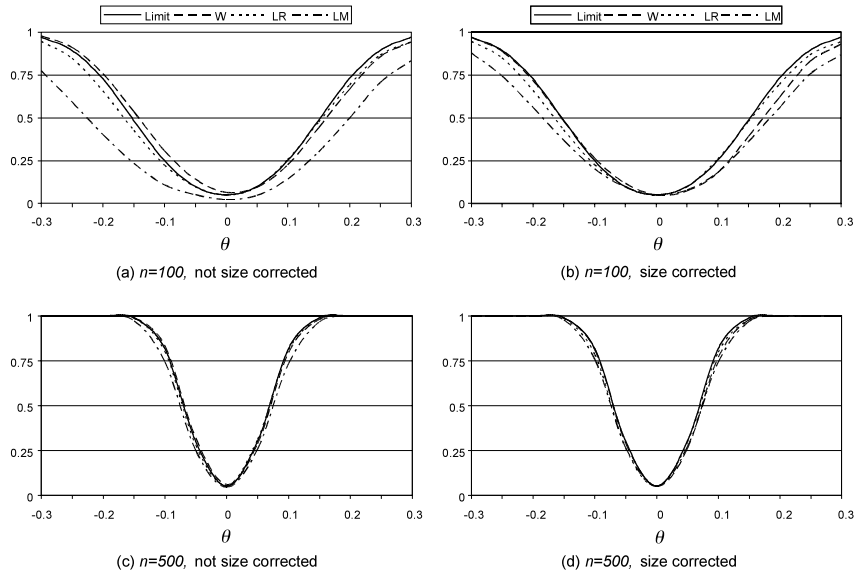

FIGURE 3. Finite sample power functions with martingale difference errors.

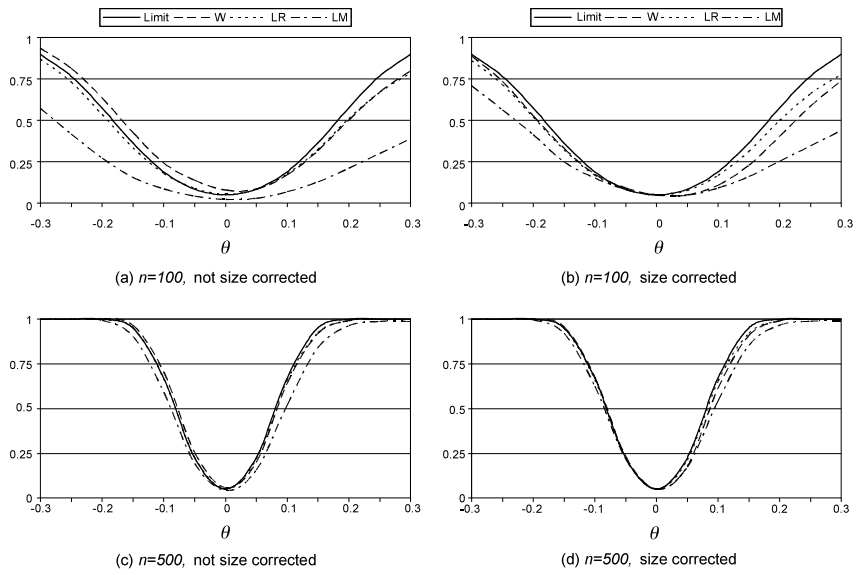

FIGURE 4. Finite sample power functions with AR(1) errors with coefficient $a = -0.5$.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

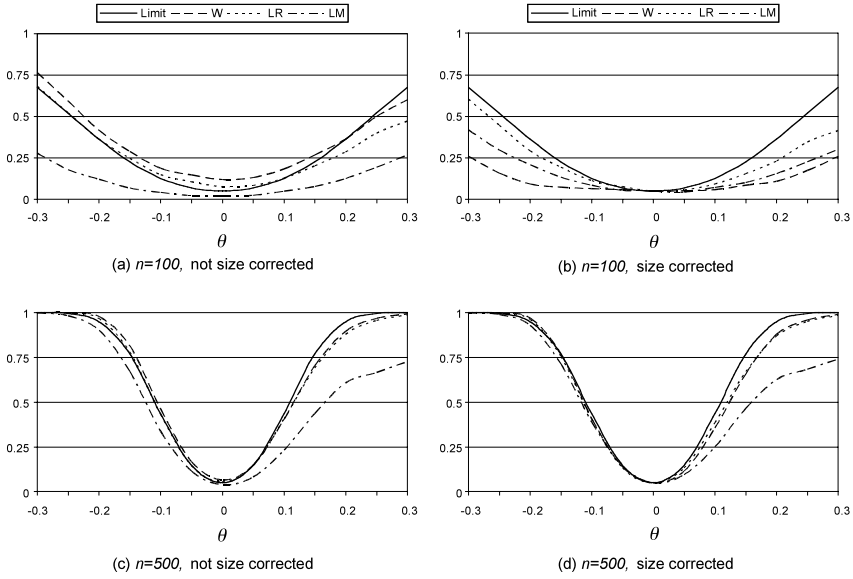

FIGURE 5. Finite sample power functions with AR(1) errors with coefficient $a = 0$.

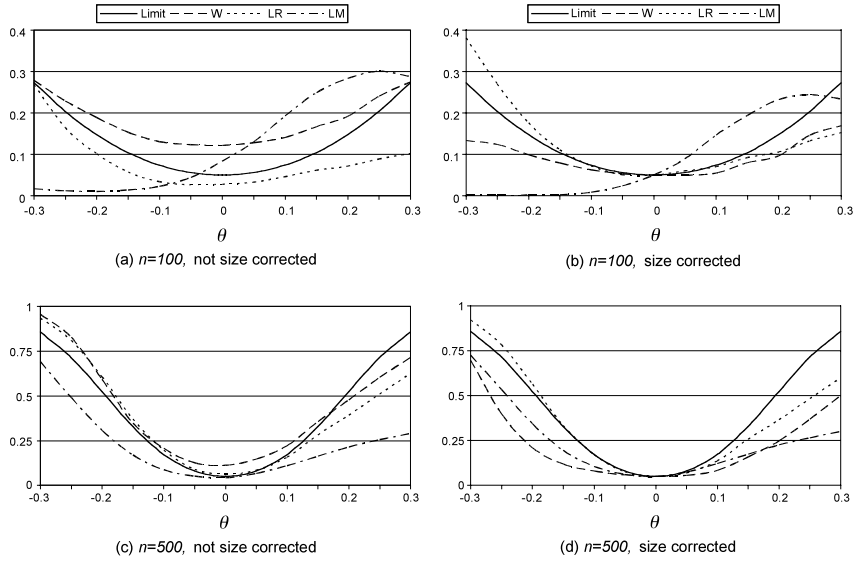

FIGURE 6. Finite sample power functions with AR(1) errors with coefficient $a = 0.5$.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

close to the asymptotic local power, except the LM test in the small sample $n = 100$, which has lower power than the LR and Wald tests.

When the errors are serially correlated the differences between the test statistics are more apparent. With negative serial correlation $a = -0.5$ (Figure 4), and with $a = 0$ (Figure 5), i.e., when there is no serial correlation in the DGP but an AR(1) is estimated, the LM test loses power compared to the LR and Wald tests, and the Wald test tends to be oversized in the small sample, which is also reflected by its very low size corrected power for $n = 100$ in Figure 5(b).

In Figure 6 the errors are positively serially correlated with $a = 0.5$. From the previous sections we know that the asymptotic local power is much lower in this case than with negative or no serial correlation. As Figure 6 shows, this is also the case for the finite sample rejection frequencies (note that the scaling along the vertical axis is different in Figures 6(a) and (b), compared with the other plots). In the small sample, $n = 100$, there are severe distortions, especially to the LM and Wald tests. The LM test completely loses power against negative alternatives, with rejection frequencies even lower than the nominal size, and the Wald test is severely oversized. When $n = 500$ the situation improves, but the LM test still has the lowest power and the Wald test is still severely oversized.

Unreported simulations (which can be obtained from the author upon request) show that, not surprisingly, the performance of the LR test (with $n = 100$) is very bad when relevant deterministic terms are left out and that the inclusion of irrelevant mean and/or trend terms decreases power against positive values of θ. This is well known from AR-based unit root tests such as the Dickey–Fuller test, where a mean (and trend) must be included if any power against nonzero mean (and trend) is desired. However, it is worth noting that, unlike in our model, the distribution of Dickey–Fuller type test statistics changes when deterministic terms are included.

Overall, the simulations show that the improvement with respect to both size and power when considering $n = 500$ instead of $n = 100$ is substantial. Thus, one would expect very good performance of the tests in financial applications, where samples are often many times larger. In such cases, the power loss resulting from the estimation of serially correlated errors would also be of less importance. It was also found that generally the LM test has lower power than the Wald and LR tests and that the Wald test is often severely oversized. We have stressed the possibility of conducting simple asymptotic inference in our model, using the chi-squared tables, and because this property is lost if size corrected critical values must be employed, this weighs heavily against the Wald test.

Even though we concentrated on the simple and well-known fractional unit root model in the present simulation study, similar relative performance is to be expected in more complicated models such as the general model in (5). Thus, the LR test is expected to outperform the Wald and LM tests with respect to both size and power also in more complicated models.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

6. CONCLUSION

We have considered likelihood inference in a wide class of potentially nonstationary univariate time series models. In such cases, inference is usually drawn in an autoregressive framework and nonstandard asymptotics apply.

In this paper we have shown that, when the estimation and testing problems are embedded in a fractional integration framework, standard asymptotics apply and desirable statistical properties of likelihood inference reemerge. In particular, there exists a local MLE that is asymptotically normal, and the classical likelihood-based tests (Wald, LR, and LM) are consistent and asymptotically chi-squared distributed under local alternatives. Under the additional assumption of Gaussianity, the local MLE is asymptotically efficient, and the tests are locally most powerful. Furthermore, in the scalar parameter case with i.i.d. Gaussian errors, our tests achieve the asymptotic Gaussian power envelope of all invariant and unbiased tests; i.e., they are asymptotically uniformly most powerful (against local alternatives) among all invariant and unbiased tests.

The Monte Carlo study shows that with sample sizes typical for macroeconomic time series the tests perform reasonably well, and with larger sample sizes such as those usually found in finance applications the performance of the tests is very good and their rejection frequencies very close to the asymptotic local power. In our Monte Carlo study the LR test dominates with respect to both size and power in finite samples. The LR test also has attractive computational features when serially correlated errors are allowed for, because it avoids a quite cumbersome calculation of covariance matrices.

The results derived in this paper could also be applied to the problem of testing for fractional cointegration when the cointegrating vector is known a priori, e.g., from economic theory. When the cointegrating vector must be estimated the results in this paper no longer apply. This presents an interesting avenue for further research which is currently under investigation by the author (see, e.g., Nielsen, 2003).

REFERENCES

Agiakloglou, C. & P. Newbold (1994) Lagrange multiplier tests for fractional difference. *Journal of Time Series Analysis* 15, 253–262.

Bierens, H.J. (2001) Complex unit roots and business cycles: Are they real? *Econometric Theory* 17, 962–983.

Brown, B.M. (1971) Martingale central limit theorems. *Annals of Mathematical Statistics* 42, 59–66.

Chung, C.F. (1996) Estimating a generalized long memory process. *Journal of Econometrics* 73, 237–259.

Dickey, D.A. & W.A. Fuller (1979) Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–431.

Doornik, J.A. (2001) *Ox: An Object-Oriented Matrix Language*, 4th ed., Timberlake Consultants Press.

Doornik, J.A. & M. Ooms (2001) A Package for Estimating, Forecasting, and Simulating Arfima Models: Arfima Package 1.01 for Ox. Working Paper, Nuffield College, Oxford.

Elliott, G., T.J. Rothenberg, & J.H. Stock (1996) Efficient tests for an autoregressive unit root. *Econometrica* 64, 813–836.
Engle, R.F. (1984) Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In Z. Griliches & M.D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, pp. 775–826. North-Holland.
Gil-Alana, L.A. (2001) Testing stochastic cycles in macroeconomic time series. *Journal of Time Series Analysis* 22, 411–430.
Gray, H., N. Zhang, & W.A. Woodward (1989) On generalized fractional processes. *Journal of Time Series Analysis* 10, 233–257.
Hall, P. & C.C. Heyde (1980) *Martingale Limit Theory and Its Application*. Academic Press.
Hylleberg, S., R.F. Engle, C.W.J. Granger, & B.S. Yoo (1990) Seasonal integration and cointegration. *Journal of Econometrics* 44, 215–238.
Lehmann, E.L. (1986) *Testing Statistical Hypotheses*, 2nd ed., Springer-Verlag.
Ling, S. & W.K. Li (1997) Fractional ARIMA-GARCH time series models. *Journal of the American Statistical Association* 92, 1184–1194.
Ling, S. & W.K. Li (2001) Asymptotic inference for nonstationary fractionally integrated autoregressive moving-average models. *Econometric Theory* 17, 738–764.
Nielsen, M.Ø. (2003) Optimal residual based tests for fractional cointegration and exchange rate dynamics. *Journal of Business and Economic Statistics*, forthcoming.
Phillips, P.C.B. (1987) Time series regression with a unit root. *Econometrica* 55, 277–301.
Phillips, P.C.B. & Z. Xiao (1998) A primer on unit root testing. *Journal of Economic Surveys* 12, 423–469.
Robinson, P.M. (1991) Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regressions. *Journal of Econometrics* 47, 67–84.
Robinson, P.M. (1994) Efficient tests of nonstationary hypotheses. *Journal of the American Statistical Association* 89, 1420–1437.
Saikkonen, P. & R. Luukkonen (1993a) Point optimal tests for testing the order of differencing in ARIMA models. *Econometric Theory* 9, 343–362.
Saikkonen, P. & R. Luukkonen (1993b) 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.
Sowell, F.B. (1990) The fractional unit root distribution. *Econometrica* 58, 495–505.
Tanaka, K. (1999) The nonstationary fractional unit root. *Econometric Theory* 15, 549–582.

APPENDIX: PROOFS

Proof of Theorem 3.1. First, by noting that $\phi(L)\tilde{u}_t = e_t + (\beta - \tilde{\beta})'\tilde{x}_t$ it is immediate that the denominator in $g(\theta)$ is

$$(\sigma^2 + o_p(1)) + \frac{1}{n}\sum_{t=1}^{n}(\beta - \tilde{\beta})'\tilde{x}_t\tilde{x}_t'(\beta - \tilde{\beta}) + \frac{1}{n}\sum_{t=1}^{n}(\beta - \tilde{\beta})'\tilde{x}_t e_t \qquad \textbf{(A.1)}$$

by Assumption 3. The last two terms are asymptotically negligible because

$$E\left\| \frac{1}{n}\sum_{t=1}^{n}(\beta - \tilde{\beta})'\tilde{x}_t\tilde{x}_t'(\beta - \tilde{\beta}) \right\| = O\left(\frac{1}{n}\operatorname{tr}(D_n^{-1/2}D_nD_n^{-1/2})\right)$$

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

by Assumption 2 and (12) and

$$E\left\| \frac{1}{n}\sum_{t=1}^{n}(\beta-\tilde{\beta})'\tilde{x}_t e_t \right\|^2 = O\left(\frac{1}{n}\operatorname{tr}(D_n^{-1/2}D_n D_n^{-1/2})\right)$$

using also uncorrelatedness of $\{e_t\}$.

The numerator in $g(\theta)$ can be written as

$$\sum_{t=1}^{n}(\phi(L)\tilde{u}_t)^2 - \sum_{t=1}^{n}(\phi(L)u_t)^2 + \sum_{t=1}^{n}(\phi(L,\theta)u_t)^2 - \sum_{t=1}^{n}(\phi(L,\theta)\hat{u}_t)^2 \tag{A.2}$$

$$+ \sum_{t=1}^{n}(\phi(L)u_t)^2 - \sum_{t=1}^{n}(\phi(L,\theta)u_t)^2. \tag{A.3}$$

By the mean value theorem we have, for some $\theta^* = \theta^*(t,n)$ such that $0 \le \|\theta^*\| \le \|\theta\|$,

$$\begin{aligned}\phi(L,\theta)u_t &= \phi(L)u_t + \theta'\frac{\partial\phi(L,\theta^*)}{\partial\theta}u_t \\ &= e_t + \frac{\delta'}{\sqrt{n}}\zeta(L)e_t + \frac{\delta'}{\sqrt{n}}(\lambda(L,\theta^*)-\zeta(L))e_t,\end{aligned} \tag{A.4}$$

where the last term has mean zero and variance $O(n^{-1}\sum_{j=1}^{\infty}\|\lambda_j(\theta^*)-\zeta_j\|^2) = o(n^{-1})$ by Assumption 1(iii) and dominated convergence. As in Robinson (1994, p. 1435), it follows that

$$\phi(L,\theta)u_t = e_t + \frac{\delta'}{\sqrt{n}}\zeta(L)e_t + o_p(n^{-1/2}) \tag{A.5}$$

uniformly in t. Using (A.5) we get that (A.3) is

$$2\sum_{t=1}^{n}\frac{\delta'}{\sqrt{n}}(\zeta(L)e_t)e_t - \sum_{t=1}^{n}\frac{\delta'}{\sqrt{n}}(\zeta(L)e_t)(\zeta(L)e_t)'\frac{\delta}{\sqrt{n}} + o_p(1). \tag{A.6}$$

For a fixed $m > 0$, consider the p-vector $v_t = \sum_{j=1}^{m}\zeta_j e_{t-j}e_t$ and the $p \times p$ matrix $V_t = \sum_{j=1}^{m}\sum_{k=1}^{m}\zeta_j\zeta_k' e_{t-j}e_{t-k}$. By Assumption 3, $EV_t = \sigma^2\sum_{j=1}^{m}\zeta_j\zeta_j'$ and applying an LLN, $n^{-1}\sum_{t=1}^{n}V_t \to \sigma^2\sum_{j=1}^{m}\zeta_j\zeta_j'$ in probability. The vector sequence $\{v_t\}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}$ because v_t is $\mathcal{F}_t/\mathcal{B}$ measurable and integrable and $E(v_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{m}\zeta_j e_{t-j}E(e_t|\mathcal{F}_{t-1}) = 0$ *a.s.* for all t. Using Assumption 3, $Ev_t v_t' = E(E(v_t v_t'|\mathcal{F}_{t-1})) = \sigma^4\sum_{j=1}^{m}\zeta_j\zeta_j'$, and by application of a martingale difference CLT (e.g., Brown, 1971; Hall and Heyde, 1980, Chap. 3.2), we establish that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}v_t \to_d N\left(0,\sigma^4\sum_{j=1}^{m}\zeta_j\zeta_j'\right). \tag{A.7}$$

Because $E\|n^{-1/2}\sum_{t=1}^{n}\sum_{j=m+1}^{n-1}\zeta_j e_{t-j}e_t\|^2 = O(\sum_{j=m+1}^{\infty}\|\zeta_j\|^2)$ can be made arbitrarily small by choosing m large by (4), we can apply Bernstein's lemma (e.g., Hall and Heyde, 1980, pp. 191–192) to conclude that (A.3) converges in distribution to $\delta'(2\Psi^{1/2}Z - \Psi\delta)\sigma^2$. Because

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

$$g(\theta) = -\frac{n}{2}\ln\left[1 - \frac{1}{n}(2W(\delta) + o_p(1))\right] = W(\delta) + o_p(1)$$

we have proven the first statement of the theorem if we show that (A.2) is asymptotically negligible.

Thus, (A.2) can be written as

$$\sum_{t=1}^{n}(\beta - \tilde{\beta})'\tilde{x}_t\tilde{x}_t'(\beta - \tilde{\beta}) + \sum_{t=1}^{n}(\beta - \hat{\beta})'\hat{x}_t\hat{x}_t'(\beta - \hat{\beta}) \tag{A.8}$$

$$-2\sum_{t=1}^{n}(\beta - \tilde{\beta})'\tilde{x}_t e_t - 2\sum_{t=1}^{n}(\beta - \hat{\beta})'\hat{x}_t(e_t + o_p(n^{-1/2})) \tag{A.9}$$

by (A.5), where $\hat{x}_t = \phi(L,\theta)x_t$. Now, (A.9) is

$$2\sum_{t=1}^{n}(\hat{\beta} - \tilde{\beta})'\tilde{x}_t e_t + 2\sum_{t=1}^{n}(\beta - \hat{\beta})'(\tilde{x}_t - \hat{x}_t)e_t + o_p(1), \tag{A.10}$$

where

$$(\tilde{x}_t - \hat{x}_t)' = \frac{\delta'}{\sqrt{n}}\zeta(L)\tilde{x}_t' + o(n^{-1/2}) \tag{A.11}$$

uniformly in t by the same analysis as for u_t, and

$$\hat{\beta} = \left(\sum_{t=1}^{n}\hat{x}_t\hat{x}_t'\right)^{-1}\sum_{t=1}^{n}\hat{x}_t(e_t + O_p(n^{-1/2})) = \tilde{\beta} + O_p(1) \tag{A.12}$$

using (A.11) and Assumption 2. Now the second term of (A.10) is $2n^{-1/2} \times \sum_{t=1}^{n}(\beta - \hat{\beta})'\sum_{j=1}^{t-1}\tilde{x}_{t-j}\zeta_j'\delta e_t + o_p(1)$ and $E\|n^{-1/2}\sum_{t=1}^{n}(\beta - \hat{\beta})'\sum_{j=1}^{t-1}\tilde{x}_{t-j}\zeta_j'\delta e_t\|^2 = O(n^{-1})$ by uncorrelatedness of $\{e_t\}$, Assumption 2, (4), (12), and (A.12). The same arguments apply to the first term of (A.10) and to the terms in (A.8).

Next, we examine

$$\frac{\partial g(\theta)}{\partial\delta} = -\left(\frac{1}{n}\sum_{t=1}^{n}(\phi(L,\theta)\hat{u}_t)^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(\frac{\partial\phi(L,\theta)}{\partial\theta}\hat{u}_t\right)\phi(L,\theta)\hat{u}_t. \tag{A.13}$$

The expression in the first set of parentheses is

$$\frac{1}{n}\sum_{t=1}^{n}(\phi(L,\theta)u_t)^2 + \frac{1}{n}\sum_{t=1}^{n}(\beta - \hat{\beta})'\hat{x}_t\hat{x}_t'(\beta - \hat{\beta})$$

$$+\frac{2}{n}\sum_{t=1}^{n}(\beta - \hat{\beta})'\hat{x}_t(e_t + O_p(n^{-1/2})) = \sigma^2 + o_p(1) \tag{A.14}$$

using (A.5), $E\|n^{-1}\sum_{t=1}^{n}(\beta - \hat{\beta})'\hat{x}_t\hat{x}_t'(\beta - \hat{\beta})\| = O(n^{-1})$, $E\|n^{-1}\sum_{t=1}^{n}(\beta - \hat{\beta})'\hat{x}_t(e_t + O_p(n^{-1/2}))\| = O(n^{-2})$ as in (A.1) by Assumption 2, (12), (A.11), and (A.12).

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

Defining the function $\zeta(z,\theta) = (\partial/\partial\theta)\ln\phi(z,\theta)$, the second sum in (A.13) is

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L,\theta)\phi(L,\theta)\hat{u}_t)\phi(L,\theta)\hat{u}_t - \frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L,\theta)\phi(L,\theta)u_t)\phi(L,\theta)u_t \tag{A.15}$$

$$+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L,\theta)\phi(L,\theta)u_t)\phi(L,\theta)u_t - \frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L)e_t)\phi(L,\theta)u_t \tag{A.16}$$

$$+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L)e_t)\phi(L,\theta)u_t - \frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L)e_t)e_t \tag{A.17}$$

$$+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L)e_t)e_t, \tag{A.18}$$

where (A.18) converges in distribution to $\Psi^{1/2}Z\sigma^2$ as in (A.6). Applying (A.5) to (A.17) we see that it equals $n^{-1}\sum_{t=1}^{n}(\zeta(L)e_t)(\zeta(L)e_t)'\delta + o_p(1)$, which converges in probability to $\Psi\delta\sigma^2$ as in (A.6).

Thus, we need to show that (A.15) and (A.16) are asymptotically negligible. First, write (A.15) as

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L,\theta)\phi(L,\theta)(\hat{u}_t - u_t))\phi(L,\theta)\hat{u}_t \\
&\quad+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L,\theta)\phi(L,\theta)u_t)\phi(L,\theta)(\hat{u}_t - u_t) \\
&=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L)(\beta-\hat{\beta})'\hat{x}_t)(e_t + (\beta-\hat{\beta})'\hat{x}_t) \\
&\quad+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}((\zeta(L,\theta)-\zeta(L))(\beta-\hat{\beta})'\hat{x}_t)(e_t + (\beta-\hat{\beta})'\hat{x}_t) \\
&\quad+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\zeta(L)e_t)(\beta-\hat{\beta})'\hat{x}_t + \frac{1}{\sqrt{n}}\sum_{t=1}^{n}((\zeta(L,\theta)-\zeta(L))e_t)(\beta-\hat{\beta})'\hat{x}_t + o_p(1)
\end{aligned}$$

using (A.5). The first and third terms are $O_p(n^{-1/2})$ by (4) and the arguments applied to (A.14), and the second and fourth terms are $O_p(n^{-1/2})$ by combining the arguments applied to the first term and those applied to (A.4). Rewriting (A.16) as $n^{-1/2}\sum_{t=1}^{n}((\lambda(L,\theta)-\zeta(L))e_t)(e_t + O_p(n^{-1/2}))$ using (A.5), we note that it is asymptotically negligible by the same arguments as applied to (A.4). This establishes the second statement of the theorem.

The second derivative is

$$\begin{aligned}
\frac{\partial^2 g(\theta)}{\partial\delta\partial\delta'} &= -\left(\frac{1}{n}\sum_{t=1}^{n}(\phi(L,\theta)\hat{u}_t)^2\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\left(\frac{\partial\phi(L,\theta)}{\partial\theta}\hat{u}_t\right)\left(\frac{\partial\phi(L,\theta)}{\partial\theta}\hat{u}_t\right)' \\
&\quad-\left(\frac{1}{n}\sum_{t=1}^{n}(\phi(L,\theta)\hat{u}_t)^2\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\left(\frac{\partial^2\phi(L,\theta)}{\partial\theta\partial\theta'}\hat{u}_t\right)\phi(L,\theta)\hat{u}_t,
\end{aligned}$$

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

which is equal to

$$-\sigma^{-2}\frac{1}{n}\sum_{t=1}^{n}(\zeta(L,\theta)\phi(L,\theta)\hat{u}_t)(\zeta(L,\theta)\phi(L,\theta)\hat{u}_t)'$$

$$-\sigma^{-2}\frac{1}{n}\sum_{t=1}^{n}(\lambda(L,\theta)\phi(L)\hat{u}_t)\phi(L,\theta)\hat{u}_t$$

$$-\sigma^{-2}\frac{1}{n}\sum_{t=1}^{n}(\zeta(L,\theta)\zeta(L,\theta)\phi(L,\theta)\hat{u}_t)\phi(L,\theta)\hat{u}_t + o_p(1)$$

by (A.14). Combining the preceding arguments it can be shown that the last two terms are both $o_p(1)$ whereas the first term converges in probability to $-\Psi$. This completes the proof. ■

Proof of Theorem 3.2. By Theorem 3.1(iii) and Assumption 1, $g(\theta)$ is asymptotically a concave function of $\delta = \sqrt{n}\theta$ in $S_p(0,\|\delta\|/\sqrt{n})$, the sphere in p-dimensional Euclidean space centered at the origin with radius $\|\delta\|/\sqrt{n}$. Hence, by Theorem 3.1 and the subsequent analysis, $\hat{\delta} = \sqrt{n}\hat{\theta}_n$ is asymptotically the unique maximizer of $W(\delta)$ in $S_p(0,\|\delta\|/\sqrt{n})$, and its asymptotic distribution is given by (15) by the usual expansion. Under Gaussianity of $\{e_t\}$, (8) is the true likelihood. The limiting Fisher information is then given by

$$\lim_{n\to\infty}\frac{1}{n}E\left(-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta=0}\right) = \Psi,$$

which is the inverse of the asymptotic variance as required. ■

Proof of Theorem 3.3. Though the equivalence of the test statistics is well known in standard testing problems, we have stressed the nonstandard nature of our model, and thus we start by showing equivalence. By the mean value theorem

$$\sqrt{n}\hat{\theta}_n = \left(\frac{1}{n}\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta^*}\right)^{-1}\left(\frac{1}{\sqrt{n}}\frac{\partial l(\theta)}{\partial\theta}\bigg|_{\theta=0}\right),$$

where θ^* is an intermediate value. This implies that $W - LM \to 0$ in probability by Theorem 3.1. Similarly, by a Taylor expansion of the likelihood

$$\begin{aligned} l(0) &= l(\hat{\theta}_n) + \hat{\theta}_n'\frac{\partial l(\theta)}{\partial\theta}\bigg|_{\theta=\hat{\theta}_n} + \frac{1}{2}\hat{\theta}_n'\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta^*}\hat{\theta}_n \\ &= l(\hat{\theta}_n) + \frac{1}{2}n\hat{\theta}_n'\left(\frac{1}{n}\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta=0}\right)\hat{\theta}_n + o_p(1), \end{aligned}$$

and thus $LR - W \to 0$ in probability by Theorem 3.1(iii).

The asymptotic distribution of the test statistics follows directly from the previous theorems. Under the local alternatives (23) we set $\sqrt{n}(\hat{\theta}_n - \theta_{1n}) = \hat{\delta} - \delta \to_d \Psi^{-1/2}Z$ by Theorem 3.2. Then the Wald test is

$$W = \hat{\delta}'\Psi\hat{\delta} \to_d (\Psi^{-1/2}Z + \delta)'\Psi(\Psi^{-1/2}Z + \delta)$$

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

by Theorem 3.2. Similarly,

$$LM = \frac{\partial g(\theta)}{\partial \theta'}\left[E_0\left(\frac{\partial g(\theta)}{\partial \theta}\frac{\partial g(\theta)}{\partial \theta'}\right)\right]^{-1}\frac{\partial g(\theta)}{\partial \theta} \to_d (\Psi^{1/2}Z - \Psi\delta)'\Psi^{-1}(\Psi^{1/2}Z - \Psi\delta)$$

by Theorem 3.1(ii) and (21), and

$$LR = 2g(\hat{\theta}_n) \to_d (\Psi^{-1/2}Z + \delta)'\Psi(\Psi^{-1/2}Z + \delta)$$

by Theorems 3.1(i) and 3.2.

Under the additional assumption of Gaussianity the tests are locally most powerful because the noncentrality parameter is maximal by Theorem 3.2 and the formula for the information matrix (21). ■

Proof of Corollary 3.1. This is immediate from Theorem 3.3. ■

Proof of Theorem 3.4. Following the arguments of the previous sections and those in Tanaka (1999) and using (A.5) we find that

$$\tilde{e}_{tn} = e_t + \frac{c}{\sqrt{n}}\sum_{j=1}^{t-1}\zeta_j e_{t-j} + o_p(n^{-1/2}),$$

$$\hat{e}_{tn} = e_t + \frac{c-\delta}{\sqrt{n}}\sum_{j=1}^{t-1}\zeta_j e_{t-j} + o_p(n^{-1/2})$$

uniformly in t. Thus, the denominator of (26) normalized by n^{-1} converges to σ^2 in probability as $n \to \infty$, and the numerator

$$\begin{aligned}\sum_{t=1}^{n}(\tilde{e}_{tn}^2 - \hat{e}_{tn}^2) &= \frac{2\delta}{\sqrt{n}}\sum_{t=1}^{n}\sum_{j=1}^{t-1}\zeta_j e_{t-j}e_t + \frac{\delta(2c-\delta)}{n}\sum_{t=1}^{n}\left(\sum_{j=1}^{t-1}\zeta_j e_{t-j}\right)^2 + o_p(1)\\ &= 2\delta\sqrt{\Psi}\sigma^2 Z + \delta(2c-\delta)\Psi\sigma^2 + o_p(1)\end{aligned}$$

by the same arguments as those in the proof of Theorem 3.1. As before, it can be shown that this is unaffected by the presence of the regressors and the result follows. ■

Proof of Theorem 3.5. Consider first (28), which implies that (in this context ϕ is the density function of the standard normal distribution)

$$\phi\left(\frac{C_{2,\alpha}(\delta) + \delta^2\Psi}{2\delta\sqrt{\Psi}}\right) = \phi\left(\frac{C_{1,\alpha}(\delta) + \delta^2\Psi}{2\delta\sqrt{\Psi}}\right)$$

with the nontrivial solution $C_{1,\alpha}(\delta) = -C_{2,\alpha}(\delta) - 2\delta^2\Psi$. Now determine the constants by (27):

$$\begin{aligned}1-\alpha &= P(-C_{2,\alpha}(\delta) - 2\delta^2\Psi < M(0,\delta) < C_{2,\alpha}(\delta))\\ &= P\left(-\frac{C_{2,\alpha}(\delta) + \delta^2\Psi}{2\delta\sqrt{\Psi}} < Z < \frac{C_{2,\alpha}(\delta) + \delta^2\Psi}{2\delta\sqrt{\Psi}}\right),\end{aligned}$$

where Z is a standard normal random variable. Thus, $C_{2,\alpha}(\delta)$ is the solution to

$$\Phi((C_{2,\alpha}(\delta) + \delta^2\Psi)/2\delta\sqrt{\Psi}) = 1 - \alpha/2,$$

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

i.e., $C_{2,\alpha}(\delta) = 2\delta\sqrt{\Psi}Z_{1-\alpha/2} - \delta^2\Psi$, where $Z_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

The power envelope is given by

$$\begin{aligned}\Pi(\delta) &= 1 - P(C_{1,\alpha}(\delta) < M(\delta,\delta) < C_{2,\alpha}(\delta)) \\ &= 1 - P(-2\delta\sqrt{\Psi}Z_{1-\alpha/2} - \delta^2\Psi < 2\delta\sqrt{\Psi}Z + \delta^2\Psi < 2\delta\sqrt{\Psi}Z_{1-\alpha/2} - \delta^2\Psi) \\ &= P(|Z + \delta\sqrt{\Psi}| > Z_{1-\alpha/2}) \\ &= F_{\delta^2\Psi}(\chi^2_{1,1-\alpha}),\end{aligned}$$

where the last line follows by squaring both sides of the inequality, $\chi^2_{1,1-\alpha}$ is the $100(1-\alpha)\%$ point of the χ^2_1 distribution, and $F_{\delta^2\Psi}$ is the distribution function of the $\chi^2_1(\delta^2\Psi)$ distribution. ■

Proof of Theorem 4.1. The proof proceeds along the same lines as those of Theorems 3.1 and 3.2. By the same arguments it can be shown that the results are unaffected by the presence of the regressors, so we assume here that $\{u_t\}$ is observed.

Under $\gamma = \gamma_0 + \mu/\sqrt{n}$, $\mu = (\delta', \nu')'$, we first show that

$$\text{(i)}\quad g(\gamma) \to_d W(\mu) = \frac{\mu'}{2}(2\Xi^{1/2}Z - \Xi\mu),$$

$$\text{(ii)}\quad \frac{\partial g(\gamma)}{\partial \mu} \to_d \frac{\partial W(\mu)}{\partial \mu} = \Xi^{1/2}Z - \Xi\mu,$$

$$\text{(iii)}\quad \frac{\partial^2 g(\gamma)}{\partial \mu \partial \mu'} \to_p -\Xi,$$

where Z is a $(p+q)$-dimensional standard normal random vector.

It is immediate that the denominator in $g(\gamma)$ converges in probability to σ^2 by Assumption 4. By the mean value theorem we have, for some $\gamma^* = \gamma^*(t,n)$ partitioned as $\gamma^* = (\theta^{*\prime}, \psi^{*\prime})'$ and such that $\|\gamma_0\| \le \|\gamma^*\| \le \|\gamma\|$,

$$\begin{aligned}\phi(L,\theta)c(L,\psi)u_t = \varepsilon_t &+ \frac{\mu'}{\sqrt{n}}\xi(L)\varepsilon_t + \frac{\delta'}{\sqrt{n}}(\lambda(L,\theta^*) - \zeta(L))\varepsilon_t \\ &+ \frac{\nu'}{\sqrt{n}}(\lambda_\nu(L,\psi^*) - \lambda_\nu(L,\psi_0))\varepsilon_t,\end{aligned}$$

where $\lambda_\nu(z,\psi) = (\partial \ln c(z,\psi)/\partial\psi)(c(z,\psi)/c(z,\psi_0))$ and $\lambda(z,\theta)$ is defined in Assumption 1(iii). Denoting by $\lambda_{\nu,j}(\psi)$ the coefficient on z^j in an expansion of $\lambda_\nu(z,\psi)$ in powers of z and by N a neighborhood of size $O(n^{-1/2})$ around ψ_0, $\sup_{\psi\in N}\sum_{j=0}^{\infty}\|\lambda_{\nu,j}(\psi)\|^2 < \infty$ because $a(z,\psi)$ and $b(z,\psi)$ have roots that are outside the unit circle. Thus, as in (A.5) it follows that

$$\phi(L,\theta)c(L,\psi)u_t = \varepsilon_t + \frac{\mu'}{\sqrt{n}}\xi(L)\varepsilon_t + o_p(n^{-1/2}) \qquad \textbf{(A.19)}$$

uniformly in t.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press

Hence, the numerator in $g(\gamma)$ is

$$2\sum_{t=1}^{n}\frac{\mu'}{\sqrt{n}}(\xi(L)\varepsilon_t)\varepsilon_t-\sum_{t=1}^{n}\frac{\mu'}{\sqrt{n}}(\xi(L)\varepsilon_t)(\xi(L)\varepsilon_t)'\frac{\mu}{\sqrt{n}}+o_p(1). \tag{A.20}$$

Define for a fixed $m>0$ the $(p+q)$-vector $v_t=\sum_{j=1}^{m}\xi_j\varepsilon_{t-j}\varepsilon_t$ and the $(p+q)\times(p+q)$ matrix $V_t=\sum_{j=1}^{m}\sum_{k=1}^{m}\xi_j\xi_k'\varepsilon_{t-j}\varepsilon_{t-k}$. As in the proof of Theorem 3.1, $n^{-1}\sum_{t=1}^{n}V_t\to\sigma^2\sum_{j=1}^{m}\xi_j\xi_j'$ in probability, and

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}v_t\to_d N\left(0,\sigma^4\sum_{j=1}^{m}\xi_j\xi_j'\right)$$

by application of a martingale difference CLT. Part (i) now follows by Bernstein's lemma.

To prove (ii) we notice that the first term in

$$\frac{\partial g(\gamma)}{\partial\mu}=-\left(\frac{1}{n}\sum_{t=1}^{n}(\phi(L,\theta)c(L,\psi)u_t)^2\right)^{-1}\times\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\xi(L,\gamma)\phi(L,\theta)c(L,\psi)u_t)\phi(L,\theta)c(L,\psi)u_t \tag{A.21}$$

is $(\sigma^2+o_p(1))^{-1}$ by (A.19) and write the second term in (A.21) as

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\xi(L,\gamma)\phi(L,\theta)c(L,\psi)u_t)\phi(L,\theta)c(L,\psi)u_t-\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\xi(L)\varepsilon_t)\varepsilon_t+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\xi(L)\varepsilon_t)\varepsilon_t.$$

The last term converges in distribution to $\Xi^{1/2}Z\sigma^2$ as in (A.20), and by application of (A.19) the difference of the first two terms is $n^{-1}\sum_{t=1}^{n}\mu'(\xi(L)\varepsilon_t)(\xi(L)\varepsilon_t)'\mu+o_p(1)$, which converges in probability to $\Xi\mu\sigma^2$ as in (A.20).

The result (iii) follows exactly as in the proof of Theorem 3.1.

Next, it follows as in Section 3.1 that (14) holds with δ replaced by μ and g replaced by the function in Section 4. Thus, the existence and uniqueness in $S_{p+q}(0,\|\mu\|/\sqrt{n})$ of a local MLE $\hat{\gamma}_n$ satisfying $\sqrt{n}\hat{\gamma}_n=O_p(1)$ are ensured, and its distribution is given by (36) from the usual expansion.

Efficiency follows directly from (iii), which is the Fisher information under Gaussianity of $\{\varepsilon_t\}$. ■

Proof of Corollary 4.1. Apply the partitioned matrix inverse formula to Ξ. ■

Proof of Theorem 4.2. This follows straightforwardly by applying the arguments in the proof of Theorem 3.3 to the results in Theorem 4.1 and its proof. ■

Proof of Corollary 4.2. This is immediate from Theorem 4.2.

https://doi.org/10.1017/S0266466604201050 Published online by Cambridge University Press