Econometrica, Vol. 71, No. 4 (July, 2003), 1269–1286

TESTS FOR UNIT ROOTS AND THE INITIAL CONDITION

BY ULRICH K. MÜLLER AND GRAHAM ELLIOTT[1]

The paper analyzes the impact of the initial condition on the problem of testing for unit roots. To this end, we derive a family of optimal tests that maximize a weighted average power criterion with respect to the initial condition. We then investigate the relationship of this optimal family to popular tests. We find that many unit root tests are closely related to specific members of the optimal family, but the corresponding members employ very different weightings for the initial condition. The popular Dickey-Fuller tests, for instance, put a large weight on extreme deviations of the initial observation from the deterministic component, whereas other popular tests put more weight on moderate deviations. Since the power of unit root tests varies dramatically with the initial condition, this paper explains the results of comparative power studies of unit root tests. The results allow a much deeper understanding of the merits of particular tests in specific circumstances, and a guide to choosing which statistics to use in practice.

KEYWORDS: Unit root tests, point optimal tests, weighted average power, asymptotic distributions.

1. INTRODUCTION

IN TESTING FOR A UNIT ROOT, one faces a large array of possible methods. Monte Carlo studies do not point to any dominant test. Part of the reason is that there exists no uniformly most powerful test (cf. Elliott, Rothenberg, and Stock (1996), abbreviated ERS in the following). In this paper we show analytically that treatment of the initial condition of the process is a second reason for the lack of such an ordering. The power of any unit root test will depend on the deviation of the initial observation y_0 from its modelled deterministic part—call this deviation ξ. We derive a family of optimal tests that maximize a weighted average power criterion with respect to ξ, and then relate popular unit root tests to this optimal family in an asymptotic framework. We find that all popular unit root tests are either members of the optimal family or closely related to a specific member.[2] These relationships allow us to infer the implicit weighting of ξ of the various popular tests, and we find stark differences in this regard. The popular Dickey-Fuller tests, for instance, put a large weight on extreme deviations of the initial observation from the deterministic component, whereas other popular tests that fare well in Monte Carlo studies, like tests based on weighted symmetric estimators, put more weight on moderate ξ.

Our findings have several implications. First, the initial condition is a crucial aspect of the unit root testing problem, as it profoundly affects both power and the form of

[1] Part of the paper was written during a visit of Müller at the University of California at San Diego. He thanks UCSD for its hospitality and the Swiss National Fund for generous financial support. Elliott acknowledges support in part from National Science Foundation Grant SES-01-11238. The authors are grateful for helpful comments from R. Giacomini, E. Pesavento, and two anonymous referees and the editor.

[2] Not all unit root tests have this property; however many do.

optimal tests. Second, there is little point in deriving yet another unit root test statistic as all of the popular tests are close to optimal for some weighting of ξ. Even if one comes up with an additional statistic that has a different and potentially attractive power characteristic in the ξ dimension, it seems much more compelling to use our general method for computing a test with this property directly. Third, the implicit weightings of ξ found for the popular tests explain the results of comparative power studies. Monte Carlo evidence is in general inconclusive. The careful analysis of the role of the initial condition provides a unifying and consistent interpretation for the simulation results. Fourth, we make clear that choices between statistics in practice come down to what types of initial conditions are likely for the application at hand and reveal the merits of particular tests for specific model parameterizations.

In the next section we build the basic model and discuss various methods for dealing with the nuisance parameter ξ. We then derive the family of optimal tests that maximize weighted average power over different initial conditions in both small and large samples. Section four relates commonly employed unit root tests to members of the optimal family.

2. HYPOTHESIS TESTING AND THE INITIAL CONDITION

We will consider the following general model in this paper:

$$\begin{aligned} &y_t = X_t'\beta + \mu + w_t && (t = 0, 1, \dots, T),\\ (2.1)\qquad &w_t = \rho w_{t-1} + \nu_t && (t = 1, \dots, T),\\ &w_0 = \xi, \end{aligned}$$

where X_t is a predetermined vector with no constant element, $X_0 = 0$, and μ, β, and ξ are unknown. We also assume that the regressor matrix $X = (X_0, \dots, X_T)'$ has full column rank. We are interested in distinguishing the two hypotheses $H_0 : \rho = 1$ vs. $H_1 : \rho < 1$.

This model has received a great deal of attention. Test statistics typically do not have approximate normal distributions, and much of the intuition from the stationary world as to which tests are optimal does not hold for this testing situation. Many feasible test statistics have been suggested, the most famous being Dickey and Fuller's (1979) t-test and $\hat{\rho}$-test. Monte Carlo evidence leads Pantula, Gonzalez-Farias, and Fuller (1994) to promote tests based on weighted symmetric regressions. None of these tests have known optimality properties.

Less work is concerned with the derivation of optimal tests. Dufour and King (1991) derive the point optimal test and locally best test for independent Gaussian disturbances ν_t and an independent zero-mean normal ξ for various ρ. ERS derive the family of asymptotically optimal tests against a fixed alternative when ξ is bounded in probability and for (possibly correlated) Gaussian ν_t. Rothenberg and Stock (1997) extend this to alternate distributions on the error terms. Elliott (1999) derives the family of asymptotically optimal tests for independent ν_t when ξ is drawn from its unconditional distribution under the fixed alternative.

The initial value ξ can be regarded as a fixed but unknown nuisance parameter. With $y = (y_0, y_1, \dots, y_T)'$, $w = (w_0, w_1, \dots, w_T)'$, and e a $T+1 \times 1$ vector of ones, we can write the model compactly as

$$y = X\beta + \mu e + w.$$

Note that the expectation of w is not zero, but rather $E[w] = \xi R(\rho)$, where $R(r)$ is the $T+1 \times 1$ vector $R(r) = (1, r, r^2, \dots, r^T)'$. Let $A(r)$ be the $T+1 \times T+1$ matrix with ones on its main diagonal, $-r$ on the lower diagonal, and zeros elsewhere, and let $\nu = (0, \nu_1, \dots, \nu_T)'$. Then $A(\rho)(w - \xi R(\rho)) = \nu$ and the model can be written

$$y = X\beta + \mu e + \xi R(\rho) + A(\rho)^{-1}\nu.$$

From a statistical perspective, the initial condition ξ is an additional nuisance parameter along with β and μ and the covariance matrix of ν. We are not primarily interested in its value, but we must be concerned about its impact on the data generating process in order to construct useful tests and evaluate their performance.

Under the null hypothesis ($\rho = 1$) different values of ξ induce mean shifts in the data, as $R(1) = e$. So with $\rho = 1$, ξ and μ have the exact same impact on the data generating process, and ξ and μ are not individually identified. This means that tests invariant to the mean will be numerically unaffected by the initial condition, so that ξ does not affect their size. Under the alternative hypothesis ($\rho < 1$) altering ξ amounts to adding a geometrically decaying series $\Delta\xi\rho^t$ to the data. This results in an extra difference between the null and alternative models and will affect power and the form of the optimal test for a unit root.

It is long known that the power of unit root tests depends on the initial condition in small samples; see Evans and Savin (1981, 1984), for instance, and Stock (1994, p. 2777) for additional references. Typical asymptotic analyses assume ξ to be either fixed or random but bounded in probability. An application of a Functional Central Limit Theorem (FCLT) to

$$T^{-1/2} w_{[Ts]} = T^{-1/2}\rho^{[Ts]}\xi + T^{-1/2}\sum_{l=1}^{[Ts]} \rho^{[Ts]-l}\nu_l \tag{2.2}$$

with $[\cdot]$ denoting the largest smaller integer function then suggests that ξ has no bearing on the relevant asymptotic distributions either under the null or alternative hypotheses as the first term is $o_p(1)$.

More adequate asymptotic approximations for small sample inference, when ξ is of similar magnitude to variation in the data after deterministic terms are removed, arises when the first term is $O(1)$. Useful asymptotics for the unit root testing problem require ρ to become ever closer to unity as the sample size T increases. Following the analysis of Chan and Wei (1987) and Phillips (1987b), the appropriate rate of convergence of ρ to one is achieved by setting $\rho = 1 - \gamma T^{-1}$ for a fixed γ. A stationary series in this framework will have an unconditional variance that is proportional to $(1-\rho^2)^{-1} = T(2\gamma)^{-1} + o(T)$. Taking the root of the unconditional variance as the natural scale for the initial condition, that suggests treating ξ as an $O(T^{1/2})$ variable. But with $\xi = O(T^{1/2})$, the first term in (2.2) has the same order of magnitude as the second, so that the initial condition does not vanish asymptotically.

With ξ being a relevant nuisance parameter, we require some method to allow for it in the testing procedure. A 'plug-in' approach that substitutes ξ with an estimator $\hat{\xi}$ in a procedure that is optimal for a specific ξ fails to provide an optimal test because ξ cannot be estimated with sufficient precision to leave power unaffected by the substitution. Alternatively one might consider tests that are invariant to ξ (as is usually done with deterministic terms). But with ρ unknown, the relevant group of transformations

$$y \to y + xR(r) \qquad \forall x \; \forall r < 1 \tag{2.3}$$

is so large that a requirement of invariance with respect to this group yields tests with trivial power. On a conceptual level the application of invariance is inappropriate for ξ

because the form of the induced deterministic $R(\rho)$ depends on ρ and thus on the upheld hypothesis.

We hence derive tests that maximize weighted average power over various values of ξ, where the weight function is a prespecified distribution function. Since tests that are invariant to the mean are unaffected by different values of ξ under the null, we only need to specify the weight function under the alternative hypothesis. In this respect, the situation here is very similar to Andrews and Ploberger's (1994) analysis of optimal asymptotic tests for the general testing problem when a nuisance parameter is present only under the alternative.[3]

3. A FAMILY OF OPTIMAL TESTS

3.1. *Small Sample Analysis*

In this section we develop optimal test statistics for the unit root testing problem and derive their asymptotic distribution. We make the following assumption concerning the generation of the disturbances ν_t. The stationarity assumption is only required for the asymptotic optimality below.

CONDITION 1: *The stationary sequence $\{\nu_t\}$ has a strictly positive spectral density function $f_\nu(\cdot)$; it has a moving average representation $\nu_t = \sum_{s=0}^{\infty} \delta_s \varepsilon_{t-s}$ where the ε_t are independent standard normal random variables and $\sum_{s=0}^{\infty} s|\delta_s| < \infty$.*

With the distribution of the stochastic element ν_t specified to be normal, we would like to apply the Neyman-Pearson Lemma to derive an optimal test statistic. But three problems arise: (i) β and μ are unknown, (ii) the alternative is composite, and (iii) there is an additional nuisance parameter ξ, that is individually identified only under the alternative.

To deal with the first problem, we will restrict attention to tests that are invariant to the group of transformations

$$y \to y + Zb \qquad \forall b, \tag{3.1}$$

where $Z = (e, X)$, i.e. the requirement that a test statistic $S(y)$ has the property $S(y + Zb) = S(y)$ for all b. This has been the dominant strategy in the unit root literature for the treatment of the unknown β and μ, and we will follow this approach. As already noted above, invariance to the mean also makes the test statistic automatically independent of ξ under the null of $\rho = 1$.

The composite nature of the alternative is indeed a problem for unit root testing, as there does not exist a uniformly most powerful test, even asymptotically (cf. ERS). Dufour and King (1991) have derived small sample point optimal tests that maximize power at a specific alternative $\rho = r < 1$, and ERS have extended these results in a local-to-unity asymptotic framework. We will follow this approach.

In order to deal with ξ, we derive tests that maximize a weighted average power criterion. Specifically, let $F(\xi)$ be a probability measure on the real line. We will refer to a (possibly randomized) test $\varphi_0(y; r, F)$ as an optimal test if for a given significance level α_0, $\varphi_0(y; r, F)$ maximizes weighted average power at the alternative $\rho = r < 1$,

$$\int_{-\infty}^{\infty} P(\varphi(y) \text{ rejects} | \rho = r, \xi = x)\, dF(x) \tag{3.2}$$

[3] We cannot directly draw on their result, however, since several of their assumptions are not satisfied for the testing problem here.

over all tests $\varphi(y)$ of size α_0. F may be seen as representing the importance a researcher attaches to the test being able to distinguish the two hypotheses for various values of ξ. In this perspective, the weighting F is a device to derive tests with a certain power characteristic as a function of ξ.

This treatment of a nuisance parameter is very similar to the approach of Andrews and Ploberger (1994) for the general testing problem where a nuisance parameter is present only under the alternative. Their analysis would suggest a second averaging over various values of ρ. In this paper, we simplify the following derivations by sticking to the formulation (3.2) and by considering only the cumulative distribution function of zero mean normals for F (see Müller (2002) for a more general treatment). These simplifications lead to a class of optimal tests that are easier to interpret while being general enough to successfully relate existing unit root tests to specific members of the class.

The following theorem provides an optimal unit root test for general X with our chosen weighting function F for ξ. We measure the variance of the normal weighting function by multiples (denoted by k) of the unconditional variance of w_t when $\rho = r < 1$, which is given by $v_0(r) = \text{var}[\sum_{j=0}^{\infty} r^j \nu_{-j}]$.[4] Choosing k larger thus gives greater weight to larger $|\xi|$.

THEOREM 1: *Consider the data generating process* (2.1) *under Condition* 1 *where the autocovariances* $\gamma(j) = E[\nu_t \nu_{t-j}]$ *are known for all* j. *Then the test of* $H_0 : \rho = 1$ *against* $H_1 : \rho = r < 1$ *that is invariant to the transformations* (3.1) *and maximizes* (3.2) *with* F *being the cumulative density function of a zero mean normal with variance* $kv_0(r)$ *rejects for small values of the statistic*

$$Q(r,k) = y'(G_1 - G_0)y$$

where $G_i = \Sigma_i^- - \Sigma_i^- Z(Z'\Sigma_i^- Z)^{-1} Z'\Sigma_i^-$, $\widetilde{V}$ *is the covariance matrix of the last* T *elements of* ν, $\Sigma_0^- = A(1)'\text{diag}(1, \widetilde{V}^{-1})A(1)$, *and for* $k > 0$, $\Sigma_1^- = A(r)'\text{diag}(kv_0(r), \widetilde{V})^{-1}A(r)$, *whereas for* $k = 0$,

$$\Sigma_1^- = A(r)' \begin{pmatrix} v_0(r)^{-1} + (r-1)^2 \tilde{e}'\widetilde{V}^{-1}\tilde{e} & (r-1)\tilde{e}'\widetilde{V}^{-1} \\ (r-1)\widetilde{V}^{-1}\tilde{e} & \widetilde{V}^{-1} \end{pmatrix} A(r)$$

with $\tilde{e}$ *a* $T \times 1$ *vector of ones.*

An assumption of known autocovariances of ν_t is, of course, unlikely to be met in practice. But $Q(r,k)$ will serve as a useful benchmark to evaluate the performance of popular unit root tests, and it will be established below that standard tests are asymptotically optimal (or close to optimal) even without the knowledge of the correlation structure of ν_t. Note that the family of tests $Q(r,k)$ contains the optimal tests considered in ERS and Elliott (1999) with $k = 0$ and $k = 1$, respectively.

As discussed above, the weighting function F may be seen as a simple device to construct a family of optimal tests $\varphi_0(y; r, F)$ with a different power characteristic in the ξ dimension, where ξ is regarded as a fixed nuisance parameter. Alternatively, one might

[4] $v_0(r)$ is necessarily finite because

$$v_0(r) = (1-r^2)^{-1}\left(\gamma(0) + 2\sum_{j=1}^{\infty} r^j \gamma(j)\right) \leq 2(1-r^2)^{-1}\sum_{j=0}^{\infty}|\gamma(j)|$$

and under Condition 1 the sequence $\gamma(j)$ is absolutely summable.

interpret the result in a Bayesian manner: With ξ random, Theorem 1 has the additional interpretation as providing a test statistic that optimally (subject to the invariance restriction) discriminates $H_0 : \rho = 1$ with arbitrary ξ against $H_1 : \rho = r$ and $\xi \sim N(0, kv_0(r))$ independent of ν.

Dufour and King's (1991) point optimal invariant statistics are closely related to $Q(r, k)$. They consider the special case where ν_t are i.i.d. Gaussian with variance σ^2, but impose invariance to the larger group of transformations of the form $y \to cy + Zb$ for any nonzero c and all vectors b. The additional invariance to scale makes the resulting tests independent of σ^2. We focus in this paper on an asymptotic analysis, and since σ^2 can be estimated consistently, the formulation of Dufour and King (1991) and $Q(r, k)$ lead to the same asymptotic power functions.

When the disturbances ν_t are independent, then the Bayesian interpretation implies that $Q(r, 1)$ is the most powerful invariant test of the unit root hypothesis against the strictly stationary alternative $\rho = r$, i.e. when ξ stems from the unconditional distribution $N(0, v_0(r))$. When ν_t is stationary but autocorrelated, however, a random ξ that makes w_t stationary under the alternative cannot be stochastically independent of ν_t. The optimal test statistic in this case is hence not a member of the family $Q(r, k)$. The following theorem provides the optimal test for this case.

THEOREM 2: *Consider the data generating process* (2.1) *under Condition* 1 *where the autocovariances $\gamma(j) = E[\nu_t \nu_{t-j}]$ are known for all j. The statistic that optimally tests $H_0 : \rho = 1$ against $H_1 : \rho = r$ and $\xi = \sum_{s=0}^{\infty} r^s \nu_{-s}$ that is invariant to the transformations* (3.1) *and rejects for small values is*

$$\overline{Q}(r) = y'(J_1 - J_0)y,$$

where

$$J_0 = G_0,$$
$$J_1 = \Omega_1^{-1} - \Omega_1^{-1} Z(Z'\Omega_1^{-1}Z)^{-1} Z'\Omega_1^{-1},$$
$$A(r)\Omega_1 A(r)' = \overline{V},$$

and

$$\overline{V} = \begin{pmatrix} v_0(r) & \tilde{\eta}' \\ \tilde{\eta} & \tilde{V} \end{pmatrix},$$

and the $T \times 1$ vector $\tilde{\eta}$ is given by $\tilde{\eta} = [\eta_t] = [\sum_{j=0}^{\infty} r^j \gamma(t + j)]$.

Whilst in small samples there is a distinction between $\overline{Q}(r)$ and $Q(r, 1)$, Theorem 3 below shows that they share the same asymptotic distribution.

3.2. *Asymptotic Analysis*

The following asymptotics are developed in the local-to-unity framework, i.e. we investigate the limiting destribution of the test statistics as the sample size T goes to infinity and $\gamma = T(1 - \rho) \geq 0$ is a fixed constant. The point alternative r against which the family of tests $Q(r, k)$ is optimal is treated accordingly as $r = 1 - gT^{-1}$ (we use γ for the true value and g for a general value). In the asymptotic analysis we measure the magnitude of the initial condition ξ in terms of the square root of the unconditional variance of a stationary process ($\gamma > 0$) with $\rho = 1 - \gamma T^{-1}$, which is $v_0(1-\gamma T^{-1})^{1/2} = \omega T^{1/2}(2\gamma)^{-1/2} + o(T^{1/2})$, where ω^2 is the 'long-run' variance of ν_t, $\omega^2 = 2\pi f_\nu(0)$. Define α implicitly as the scaled version of the initial condition, $\xi = \alpha\omega T^{1/2}(2\gamma)^{-1/2}$, so that $\xi = O(T^{1/2})$ matters asymptotically. A value of $\alpha = 1$ then generates relevant asymptotics for a finite sample where the initial condition is equal to one standard deviation of the unconditional distribution of y_t when $\rho < 1$.

In this framework, it is straightforward to show by means of an adequate FCLT and the Continuous Mapping Theorem (CMT) that

$$\text{(3.3)} \qquad T^{-1/2}(w_{[Ts]} - w_0) \Rightarrow \omega M(s)$$
$$= \begin{cases} \omega W(s) & \text{for } \gamma = 0, \\ \omega\alpha(e^{-\gamma s} - 1)(2\gamma)^{-1/2} + \omega \int_0^s e^{-\gamma(s-\lambda)} dW(\lambda) & \text{else} \end{cases}$$

where '$\Rightarrow$' denotes weak convergence of the underlying probability measures and $W(\cdot)$ is a standard Brownian motion. Note that $M(s)$ is continuous in γ at $\gamma = 0$ (cf. Elliott (1999)).

The subsequent derivations focus on the two most popular cases for the deterministic component: the mean only case without X, which will be denoted by a superscript μ, and the mean and trend case $X = \tau \equiv (0, 1, \ldots, T)'$, denoted with a superscript τ. See the longer working paper version of this contribution for the analysis of more general deterministics.

For the time trend case, it is useful to write the asymptotic distributions in terms of the projection of $M(s)$ off s, denoted $M^\tau(s)$, i.e. $M^\tau(s) = M(s) - 3s\int \lambda M(\lambda)\,d\lambda$ (for notational convenience, the limits of integration are understood to be zero and one, if not stated otherwise). In order to simplify notation, we reparameterize the families of optimal test statistics, denoted with a subscript a, which are given by $Q_a(g, k) = Q(1 - gT^{-1}, k)$ and $\overline{Q}_a(g) = \overline{Q}(1 - gT^{-1})$. The following theorem states the asymptotic distributions of $Q_a^i(g, k)$ and $\overline{Q}_a^i(g)$ in terms of $M^\mu(s) \equiv M(s)$ and $M^\tau(s)$.

THEOREM 3: *Under Condition* 1 *and with* $T(1-\rho) = \gamma \geq 0$, *for* $i = \mu, \tau$,

(i) $Q_a^i(g,k) \Rightarrow q_0^i + q_1^i M^i(1)^2 + q_2^i(\int M^i(s)\,ds)^2 + q_3^i M^i(1)\int M^i(s)\,ds + q_4^i \int M^i(s)^2\,ds$

where $q_0^\mu = -g$, $q_1^\mu = g - gk/(2+gk)$, $q_2^\mu = -g^3k/(2+gk)$, $q_3^\mu = -2g^2k/(2+gk)$, $q_4^\mu = g^2$, *and* $q_0^\tau = -g$, $q_1^\tau = (8g^2 + 8g^3 - 3g^3k + g^4k)/(24 + 24g + 8g^2 + g^3k)$, $q_2^\tau = -4g^3(3 + 3g + g^2)k/(24 + 24g + 8g^2 + g^3k)$, $q_3^\tau = 4g^3(3+g)k/(24 + 24g + 8g^2 + g^3k)$, $q_4^\tau = g^2$;

(ii) $\overline{Q}_a^i(g)$ *has the same asymptotic distribution as* $Q_a^i(g, 1)$.

The statistics of Theorem 3 were all computed with the knowledge of the covariance matrix of ν. But their asymptotic distributions do not depend on the specific form of the autocorrelations of ν_t. This—maybe surprising—result has already been established by ERS for the statistic $Q_a(g, 0)$ in our notation. It carries over to more general assumptions concerning the initial condition, as well as to the optimal statistic against the stationary model $\overline{Q}_a(g)$. The result implies that it is impossible to exploit autocorrelations in ν_t to devise unit root tests that have higher asymptotic local power than optimal tests for independent ν_t. Furthermore, since the optimal statistics have an asymptotic distribution that is a continuous function of $M(\cdot)$ and ω is consistently estimable, one can build on the results of Stock (2000) to derive feasible statistics that have the same asymptotic distribution (and are hence asymptotically optimal) that do not require knowledge of the autocovariances of ν_t. Such statistics attain the same asymptotic power as the optimal tests as long as (3.3) holds, that is, under much more general conditions than Condition 1.

Figure 1 depicts the asymptotic power of $Q_a^{\mu}(7, k)$ and $Q_a^{\tau}(13.5, k)$ with $k = 0, 1, \infty$ and $\gamma = 5, 10, 15, 20, 25$ as a function of α (the values for g are those suggested by ERS). All power curves in this paper are for a level of 5%. For large enough $|\alpha|$, the power of the optimal tests with $k = 0$ and $k = 1$ drops to zero. The tests with $k = 0$ achieve the maximal power at $\alpha = 0$ but their power drops to zero for $|\alpha| > 2$ for all considered values of γ. The tests with $k = 1$ have an asymptotic power that is lower for $\alpha = 0$ but decreases in $|\alpha|$ at a considerably slower rate. The optimal tests $Q_a^i(\cdot, \infty)$ with an extreme weighting of large $|\alpha|$ have power that increases in $|\alpha|$, and have very low power for $|\alpha| < 2$. The figures hence demonstrate the quantitative importance of the power trade-off of optimal unit root tests with respect to the weighting of the initial condition. The results are not sensitive to the choice of g, at least for moderate k.

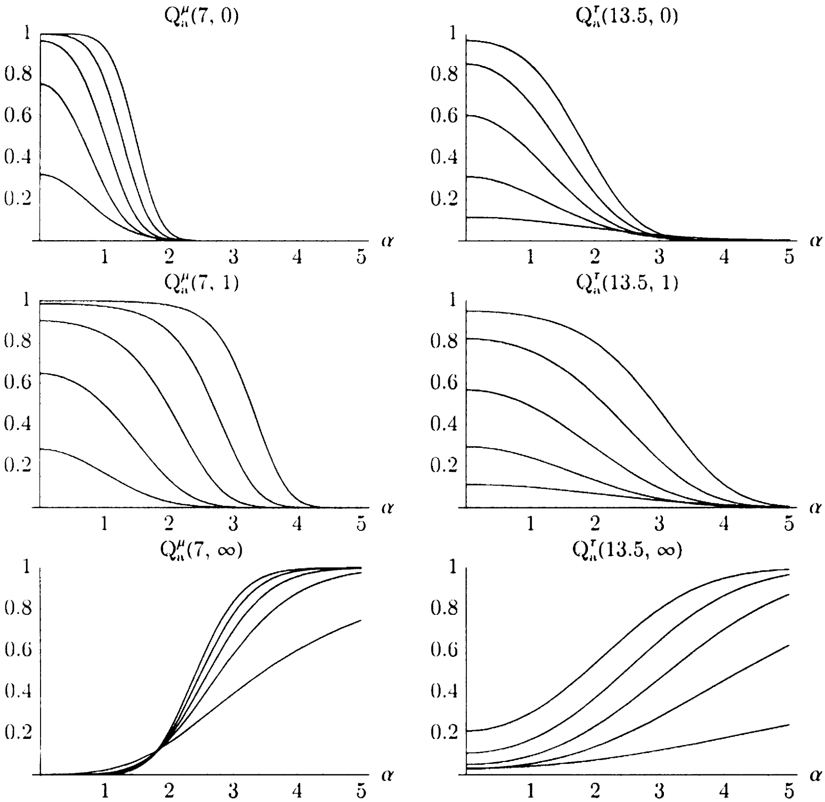

FIGURE 1.—Asymptotic power as a function of α for $\gamma = 5, 10, 15, 20$, and 25.

4. RELATION OF OPTIMAL TESTS TO SOME POPULAR UNIT ROOT TESTS

In this section we explore the relationship of some popular unit root tests to the family of optimal tests $Q_a^i(g,k)$. We compare the asymptotic distribution of $Q_a^i(g,k)$ in Theorem 3 on the one hand with the asymptotic distributions of the popular tests on the other. This is an interesting exercise because a close correspondence implies near optimality of the popular test, and the value of k of the corresponding optimal test measures the weighting of the initial condition that is implicitly employed in the popular test.

Following Stock (2000), we classify unit root test statistics by their asymptotic distributions, written as some function $h : C[0,1] \longmapsto \mathbb{R}$ of $M(\cdot)$, where $C[0,1]$ is the space of square integrable continuous functions on the unit interval. Various tests lead to a multitude of functions h, and many other tests could easily be devised. However, for most classes of tests and resulting functions h, nothing is known about their optimality.

Table I shows the classes of tests considered in this section. Members of the $\hat{\rho}^{\mathrm{DF}}$-class and $\hat{\tau}^{\mathrm{DF}}$-class include the statistic suggested by Dickey and Fuller (1979) as well as those of Phillips (1987a) and Phillips and Perron (1988), members of the N-class and R-class include the (appropriately scaled) N_1, N_2 and R_1, R_2 statistics of Bhargava (1986) as well as the t-statistics suggested by Schmidt and Phillips (1992) and Schmidt and Lee (1991), members of the LB-class include the (appropriately scaled) locally best invariant test for the mean case as derived in Dufour and King (1991) and the locally best unbiased invariant test for the trend case as derived by Nabeya and Tanaka (1990), members of the $\hat{\rho}^{\mathrm{DFGLS}}$-class and $\hat{\tau}^{\mathrm{DFGLS}}$-class, indexed by a positive parameter $\bar{c}$ in the trend case, include the statistics proposed in ERS and members of the $\hat{\tau}^{\mathrm{WS}}$-class include the weighted symmetric estimator of Pantula, Gonzalez-Farias, and Fuller (1994), where appropriate corrections for correlated disturbances are employed for all test statistics (see Stock (1994) for details regarding this correction).

TABLE I

CLASSES OF UNIT ROOT TESTS[a]

#	Class	Asymptotic Distribution
1	N	$[\int M^{i,N}(s)^2 ds]^{-1}$
2	R	$[\int M^{i,R}(s)^2 ds]^{-1}$
3	LB	$\begin{cases} M(1)^2 & \text{mean case} \\ \int M^{\tau,N}(s)^2 ds & \text{trend case} \end{cases}$
4	$\hat{\rho}^{\mathrm{DFGLS}}(\bar{c})$	$\frac{M^{i,P}(1)^2 - M^{i,P}(0)^2 - 1}{2\int M^{i,P}(s)^2 ds}$
5	$\hat{\rho}^{\mathrm{DF}}$	$\frac{M^{i,\mathrm{OLS}}(1)^2 - M^{i,\mathrm{OLS}}(0)^2 - 1}{2\int M^{i,\mathrm{OLS}}(s)^2 ds}$
6	$\hat{\tau}^{\mathrm{DFGLS}}(\bar{c})$	$\frac{M^{i,P}(1)^2 - M^{i,P}(0)^2 - 1}{2\sqrt{\int M^{i,P}(s)^2 ds}}$
7	$\hat{\tau}^{\mathrm{WS}}$	$\frac{M^{i,\mathrm{OLS}}(1)^2 + M^{i,\mathrm{OLS}}(0)^2 - 1 - 2\int M^{i,\mathrm{OLS}}(s)^2 ds}{2\sqrt{\int M^{i,\mathrm{OLS}}(s)^2 ds}}$
8	$\hat{\tau}^{\mathrm{DF}}$	$\frac{M^{i,\mathrm{OLS}}(1)^2 - M^{i,\mathrm{OLS}}(0)^2 - 1}{2\sqrt{\int M^{i,\mathrm{OLS}}(s)^2 ds}}$

[a] $M^{\mu,\mathrm{OLS}}(s) = M(s) - \int M(\lambda)d\lambda$, $M^{\tau,\mathrm{OLS}}(s) = M^{\tau}(s) - 4\int M^{\tau}(\lambda)d\lambda + 6s\int M^{\tau}(\lambda)d\lambda$, $M^{\mu,N}(s) = M(s)$, $M^{\tau,N}(s) = M^{\tau}(s) - sM^{\tau}(1)$, $M^{\mu,R}(s) = M^{\mu,\mathrm{OLS}}(s)$, $M^{\tau,R}(s) = M^{\tau}(s) - (s - \frac{1}{2})M^{\tau}(1) - \int M^{\tau}(\lambda)d\lambda$, $M^{\mu,P}(s) = M(s)$, and $M^{\tau,P}(s) = M^{\tau}(s) - s(\bar{c}+1)(\frac{1}{3}\bar{c}^2 + \bar{c} + 1)^{-1}M^{\tau}(1)$, $\bar{c} > 0$.

Noting that for a positive random variable B, $P(A/B < \text{cv}) = P(A < \text{cv } B)$, it is possible to show that a test based on statistics 1–5 in Table I with critical value cv is asymptotically equivalent to tests based on a statistic S_φ^i with asymptotic distribution

$$\varphi_0^i + \varphi_1^i M^i(1)^2 + \varphi_2^i \Big(\int M^i(s)\,ds\Big)^2 + \varphi_3^i M^i(1) \int M^i(s)\,ds + \varphi_4^i \int M^i(s)^2\,ds \tag{4.1}$$

that reject for negative values, where some of the statistic specific weights φ_j^i depend on the critical value cv. Note that (4.1) is of the same form as the asymptotic distribution of $Q_a^i(g,k)$. If there exists $l_0 > 0$ such that

$$(\varphi_1^i, \varphi_2^i, \varphi_3^i, \varphi_4^i) = l_0(q_1^i, q_2^i, q_3^i, q_4^i) \tag{4.2}$$

where q_j^i are the (nonlinear) functions of g and k given in Theorem 3, then S_φ^i and the optimal test $Q_a^i(g,k)$ have the same local power. For such an asymptotic equivalence to hold, four nonlinear equations must be satisfied by the three parameters g, k, and l_0. The next theorem describes for which statistics in Table I this equation can be satisfied.

THEOREM 4: *Under the conditions of Theorem* 3 *the classes of unit root tests* 1–4 *of Table I are asymptotically equivalent to optimal tests based on $Q_a^i(g,k)$ for a particular choice of g and k*:

	mean		*mean and trend*	
	g	k	g	k
R	∞	$\neq 0$	$\to 0$	$2/g$
N	∞	0	$\to 0$	*arbitrary constant*
LB	$\to 0$	$k < 2$	$\to 0$	*arbitrary constant*
$\hat{\rho}^{\text{DFGLS}}$	-2 cv	0	$g^{\tau,\text{DFGLS}}$	0

where the equivalence for $\hat{\rho}^{\text{DFGLS}}(\bar{c})$ in the trend case holds provided

$$g^{\tau,\text{DFGLS}} = \frac{1 - 3a + (1 - 2a - 3a^2)^{1/2}}{2a}$$

with

$$a = -\frac{\bar{c}^4 - 6\,\text{cv} - 12\bar{c}\,\text{cv} - 6\bar{c}^2\,\text{cv}}{2(3 + 3\bar{c} + \bar{c}^2)^2\,\text{cv}}$$

is real.

The locally best tests that make up the LB class were derived by Dufour and King (1991) and Nabeya and Tanaka (1990) under the assumption that the variance of ξ is a fixed number. This corresponds to the case $k = 0$, and so by construction the LB-class of tests is asymptotically equivalent to a test based on the (appropriately scaled) limit of $Q_a(g,0)$ as $g \to 0$. But Theorem 4 additionally implies that this limit is independent of k for $k < 2$. Additionally, since the asymptotic distribution of $\overline{Q}_a(g)$ is the same as that of $Q_a(g,1)$, the LB-class of tests is also asymptotically locally optimal when the initial observation is drawn from the unconditional distribution under the alternative.

The (uncorrected) R and N statistics were constructed as approximations to the locally best tests of the unit root hypothesis for independent disturbances ν_t against the stationary model ($\xi = \sum_{s=0}^{\infty} \rho^s \nu_{-s}$) and nonstationary model with $\xi = \nu_0$ in the neighborhood of $\rho = 1$, respectively. Their derivation by Sargan and Bhargava (1983) and Bhargava (1986) uses the Anderson approximation to the covariance matrices in the Gaussian densities. Interestingly, the different assumption concerning the initial condition in the derivation of N and R leads to asymptotically different approximate locally best tests, in contrast to the exact locally best tests based on LB. As already pointed out by Nabeya and Tanaka (1990), the N and R statistics generally do not—even asymptotically—correspond to the locally best test statistics when the exact densities are used. In fact, the R^μ and N^μ statistics are optimal for a ρ that is just smaller than any alternative considered in the local-to-unity framework, and a test based on R^τ is locally optimal against the alternative that ξ is Gaussian with a variance that is an order of magnitude larger than the variance of the unconditional distribution.

The $\hat{\rho}^{i,\text{DFGLS}}$-class of tests are asymptotically equivalent to a test based on $Q_a^i(g,0)$ where g depends on the level of the test. Fixing $\bar{c}$ at 13.5 in the trend case (the value suggested by ERS), we find with the 5% critical values of $\hat{\rho}^{\mu,\text{DFGLS}}$ and $\hat{\rho}^{\tau,\text{DFGLS}}$ that these tests correspond to $Q_a^\mu(16.08,0)$ and $Q_a^\tau(29.20,0)$ whereas for 1% critical values the correspondences are to $Q_a^\mu(27.39,0)$ and $Q_a^\tau(36.14,0)$. The reduction of the level therefore yields tests that are optimal for alternatives that are easier to distinguish.

The set of equations (4.2) does not have a solution for $\hat{\rho}^{\text{DF}}$, and the statistics 6–8 cannot be written in the form (4.1). Thus these statistics are not in the optimal family. But it is still insightful to identify particular values of g and k such that tests based on a class of statistics 5–8 are roughly equivalent to tests based on $Q_a(g,k)$. By (approximately) maximizing the asymptotic probability that either both tests reject or do not reject with respect to g and k under H_0, we obtain the results depicted in Table II. The column cp is the (estimated) conditional asymptotic probability that the 5% level test based on $Q_a(g,k)$ rejects given that the 5% level test based on the statistic in the first column rejects. See the Appendix for details on the selection of suitable values of g and k.

The generally large values of cp imply that the behavior of the classes of test statistics 5–8 can be mimicked very closely by specific members of the optimal family $Q_a(g,k)$, maybe with the exception of $\hat{\tau}^{\text{DF}}$ in the mean case. We corroborate these close correspondences by examining power curves for each test and the approximate optimal test—see Figure 2. For most cases the asymptotic power of the popular tests (solid lines) is hardly distinguishable from the corresponding optimal tests (dashed lines) over a wide range of values of γ and α.

The class of tests $\hat{\tau}^{\text{WS}}$ are very much comparable to tests based on $Q_a(g,1)$ for some g (since k is close to one for these tests). The implicit weighting of different α of these tests almost corresponds to the optimal weighting if the initial value is drawn from the unconditional distribution. This explains why $\hat{\tau}^{\text{WS}}$ does well in such Monte Carlo designs—see

TABLE II

CLASSES OF UNIT ROOT TESTS AND VALUES OF g AND k OF A COMPARABLE TEST BASED ON $Q_a(g,k)$

	Mean			Mean and Trend		
	cp	g	k	cp	g	k
$\hat{\tau}^{\text{WS}}$	0.983	10.8	0.988	0.991	15.9	1.01
$\hat{\tau}^{\text{DFGLS}}$	0.996	8.36	0.00	0.999	17.7	0.00
$\hat{\rho}^{\text{DF}}$	0.995	27.4	30.1	0.988	41.9	24.3
$\hat{\tau}^{\text{DF}}$	0.810	8.60	9.77	0.927	21.3	∞

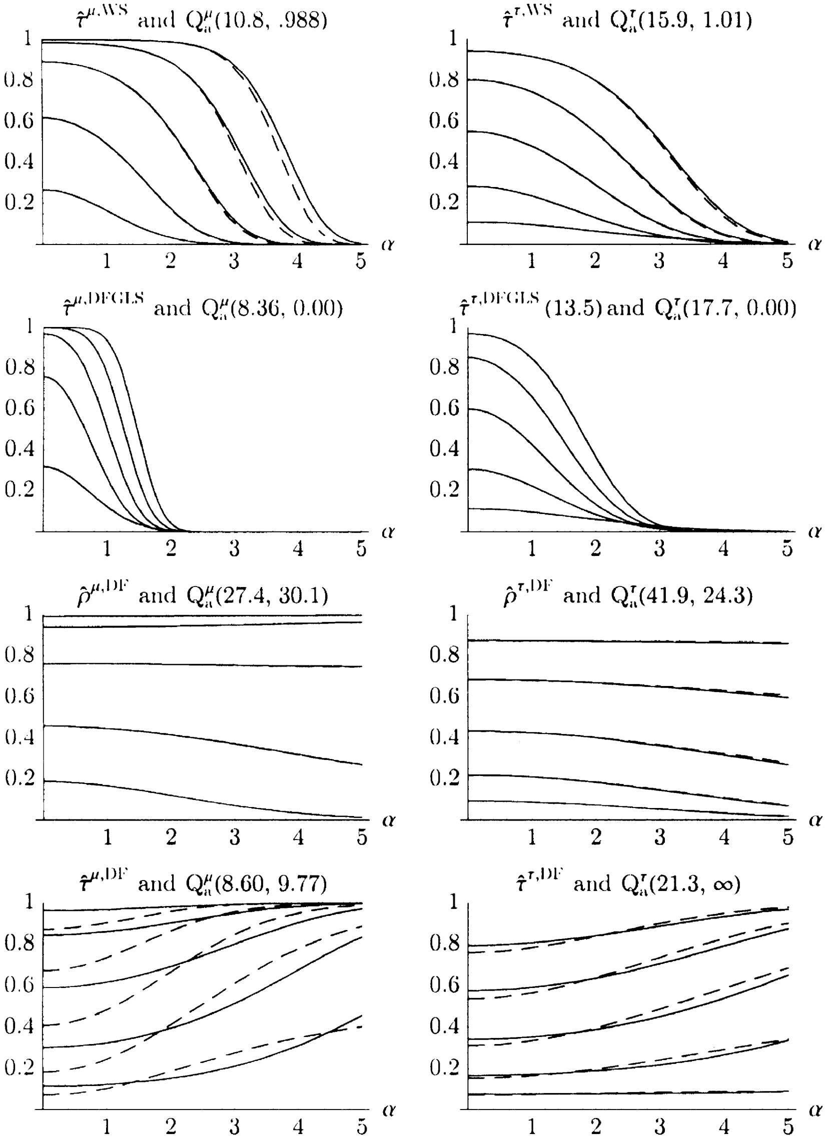

FIGURE 2.—Asymptotic power as a function of α for $\gamma = 5, 10, 15, 20$, and 25.

Pantula, Gonzalez-Farias, and Fuller (1994) and Elliott (1999). The class $\hat{\tau}^{\text{DFGLS}}$ is, as found in ERS, near optimal when α is near zero since the choice of k here is zero. Thus these results explain the available Monte Carlo evidence for this test as well. Finally, the Dickey-Fuller statistics put an extreme weighting on large $|\alpha|$, which results in asymptotic power that increases with $|\alpha|$, in stark contrast to all other considered statistics. But this increase in power for large $|\alpha|$ comes at the cost of much reduced power for small $|\alpha|$—in the mean case, for instance, the local power for $\alpha = 0$ and $\gamma = 10$ is 45.9% for $\hat{\rho}^{\mathrm{DF}}$, but 75.6% for $Q_a(7, 0)$. Given that Monte Carlo studies typically generate moderate initial conditions, it is hence not surprising that the Dickey-Fuller statistics fail to be very powerful in such set-ups.

5. CONCLUSION

In choosing a test, the choice comes down to choosing a power function. While the form of this function in terms of trade-offs over different alternatives might be debatable, certainly good tests must not have a power function that can be dominated uniformly over all alternatives. In contrast to most contributions to the unit root literature, this paper has derived a family of tests that by construction possesses this property. We related the family of optimal tests to existing tests and found that all popular unit root tests are either close or close to optimal tests. Their idiosyncratic power characteristics can be explained by different implicit treatments of the initial condition.

The near optimality of existing procedures implies that it is impossible to develop a unit root test that is uniformly better, at least within the standard assumptions on the data generating process. Continuing attempts to do so hence must be futile; there is simply no inefficiency left to exploit. Useful additions to the literature rather arise by considering the implications of a different set of assumptions for optimal procedures—see, for instance, the analysis by Rothenberg and Stock (1997) of how one might exploit nonnormality of the disturbances in the unit root testing context.

Interestingly, the near optimality holds even with respect to the Dickey-Fuller statistics. Already the first attempt at deriving a unit root testing procedure hence did not leave any 'free lunch' on the table. This paper makes plain that many ad hoc suggestions for 'better' unit root tests were in effect just trading more power at some initial condition for less power at other ones. Given the number of proposed statistics, it is perhaps not even surprising that tests that 'survived' are close to optimal.

While it is quite clear that inefficient tests should not be used, which efficient test to pick is a much more difficult question. One way to interpret our finding that efficient unit root tests have greatly varying power in the dimension of the initial condition is that knowledge about the initial condition is very informative for the problem. If a researcher is confident that reasonable initial conditions are relatively small, then it is precisely this knowledge that will enable him to generate more discriminatory power for the unit root testing problem. A useful choice then is, for instance, the tests suggested by ERS. In absence of any specific knowledge about the beginning of the series, assuming a weighting under the alternative that corresponds to the unconditional distribution of a stationary process seems like a useful starting point. This would suggest using tests based on the weighted symmetric estimator or the tests suggested by Elliott (1999). Alternatively, one could rely on the results of this paper to construct an asymptotically optimal statistic for this purpose. Given that the power of tests based on Dickey-Fuller statistics is increasing in the magnitude of the initial condition, these tests seem attractive only when there are compelling reasons why the potentially mean reverting series should start far off its equilibrium value. At any rate researchers should keep the effect of the initial condition in mind while interpreting results, and choose tests that accord to initial conditions they find sensible for their application.

Department of Economics, University of St. Gallen, Dufourstr. 48, 9000 St. Gallen, Switzerland; ulrich.mueller@unisg.ch,

and

Department of Economics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, U.S.A.; gelliott@weber.ucsd.edu

Manuscript received December, 2001; final revision received December, 2002.

APPENDIX

NOTATION: The following derivations often make use of the $(T+1)\times 1$ vector u, which is defined as $u = w - \xi e$. Vectors with tildas have T elements that correspond to the last T elements of the same vector without a tilda, i.e. $\widetilde{R} = (r, r^2, \dots, r^T)'$, $\tilde{\tau} = (1, 2, \dots, T)'$, and so forth. Similarly, $T \times T$ matrices like the covariance matrix of $\tilde{\nu}, \widetilde{V}$ or the $T \times T$ identity matrix $\widetilde{I}$ also carry a tilda. Furthermore, if A is a $T \times q$ matrix $A = [a_{tj}]$, then let A_{-1} be the $T \times q$ matrix $A_{-1} \equiv B = [b_{tj}]$, where for $1 < t \leq T$, $b_{tj} = a_{t-1,j}$ and $b_{1j} = 0$. Define the operator Δ as $\Delta A \equiv A - A_{-1}$. The limit and limit in probability as $T \to \infty$ are denoted '$\to$' and '$\overset{p}{\to}$'.

PROOF OF THEOREM 1: From the theory about invariant tests, we know that an optimal invariant procedure can always be written as a function of a maximal invariant (cf. Lehmann (1986, p. 285)). Let $M = I - Z(Z'Z)^{-1}Z'$, where I is the $(T+1)\times(T+1)$ identity matrix. Then My is such a maximal invariant. Maximizing the weighted average power criterion (3.2) is equivalent to maximizing power against the simple alternative H_1^*: the density of y is given by $\int_{-\infty}^{\infty} f(y|r, x)\,dF(x)$, where $f(y|r,x)$ is the density of y given $\rho = r$ and $\xi = x$ (cf. Andrews and Ploberger (1994)). With F the cumulative distribution function of a zero mean Gaussian with variance $kv_0(r)$, the density of $w = y - X\beta - \mu e$ under H_1^* is given by a zero mean multivariate normal with covariance matrix $\Sigma_1 = A(r)^{-1}\text{diag}(kv_0(r), \widetilde{V})A(r)'^{-1}$, so that $My = Mw \sim N(0, M\Sigma_1 M)$. Under H_0 the (singular) covariance matrix of w is given by $\Sigma_0 = A(1)^{-1}\text{diag}(0, \widetilde{V})A(1)'^{-1}$, leading to $My|H_0 \sim N(0, M\Sigma_0 M)$. Noting that the common null space of $M\Sigma_i M$ for $i = 0, 1$ is the column space of Z, the Neyman-Pearson lemma leads to optimal discrimination of the two hypotheses by the following statistic (cf. Rao and Mitra (1971, p. 206)):

$$Q(r,k) = y'M(M\Sigma_1 M)^- My - y'M(M\Sigma_0 M)^- My$$

where $(\cdot)^-$ is any generalized inverse.

We are hence left to show that the G_i are generalized inverses of $M\Sigma_i M$. Recall that a generalized inverse G of the matrix H has the property $HGH = H$. Since $G_i Z = 0$ and $Z'G_i = 0$, G_i has a row and column space no larger than the projection matrix M, so that $MG_iM = G_i$. We find

$$M\Sigma_i M G_i M \Sigma_i M = M\Sigma_i[\Sigma_i^- - \Sigma_i^- Z(Z'\Sigma_i^- Z)^{-1}Z'\Sigma_i^-]\Sigma_i M.$$

For $i = 1$ and $k > 0$, $\Sigma_1^- = \Sigma_1^{-1}$, so that we find $M\Sigma_1 M G_1 M \Sigma_1 M = M\Sigma_1 M$ immediately. For $i = 0$ or $i = 1$ and $k = 0$, we compute $\Sigma_i \Sigma_i^- \Sigma_i = \Sigma_i$, and

$$M\Sigma_i\Sigma_i^- Z = M\begin{pmatrix} 0 & 0 \\ -\tilde{e} & \widetilde{I}\end{pmatrix} Z = 0,$$

yielding $M\Sigma_i M G_i M \Sigma_i M = M\Sigma_i M$, as required. *Q.E.D.*

PROOF OF THEOREM 2: As in the proof of Theorem 1, My is a maximal invariant, and the aim is to optimally discriminate between the two multivariate normals $My|H_1 \sim N(0, M\Omega_1 M)$ and $My|H_0 \sim N(0, M\Sigma_0 M)$. The proof now follows from the same steps as for the proof of Theorem 1 above with Σ_1^- replaced by Ω_1^{-1}. *Q.E.D.*

LEMMA 1: *Let* $\widetilde{B} = (\tilde{e}, T^{-1}\tilde{\tau}_{-1}, \widetilde{R}(\rho))$. *Then under the conditions of Theorem* 3,
(i) $T^{-1}\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\widetilde{B} \to 0$,
(ii) $T^{-3/2}\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\tilde{u}_{-1} \overset{p}{\to} 0$,
(iii) $T^{-1/2}\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\Delta\tilde{u} \overset{p}{\to} 0$,
(iv) $T^{-2}\tilde{u}'_{-1}(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\tilde{u}_{-1} \overset{p}{\to} 0$,
(v) $2T^{-1}\Delta\tilde{u}'\widetilde{V}^{-1}\tilde{u}_{-1} + 1 - \omega^{-2}[2T^{-1}\Delta\tilde{u}'\tilde{u}_{-1} + \gamma(0)] \overset{p}{\to} 0$.

PROOF: The proof of the theorem draws heavily on the Appendix of ERS. For any matrix K, let $|K| = \mathrm{tr}^{1/2}[K'K]$ and note that for real conformable matrices K and L, $|\mathrm{tr}(KL)| \le |K||L|$, and $|KL| \le |K|\mathrm{r}(L) \le |K||L|$, where $\mathrm{r}(L)$ is the largest characteristic root of L. Define the $T \times T$ matrices $\widetilde{\Lambda} \equiv \omega\widetilde{V}^{-1/2} - \omega^{-1}\widetilde{V}^{1/2}$, $\widetilde{\Psi}$ the Toeplitz matrix of $\rho(j)$, the Fourier coefficients of $[2\pi f_v(\cdot)]^{-1}$ and $\widetilde{F}^r = [F^r_{ij}]$ with $F^r_{ij} = r^{i-j}$ if $i > j$ and 0 otherwise. Furthermore, let $\hat{u} = \tilde{w} - \xi\widetilde{R}(\rho)$, such that $\hat{u}$ is the disturbance vector purged of the influence of ξ. ERS show in their Appendix that under Condition 1, $0 < \omega < \infty$, $\omega^2 = \sum_{j=-\infty}^{\infty}\gamma(j)$, $\omega^{-2} = \sum_{j=-\infty}^{\infty}\rho(j)$, both $\gamma(j)$ and $\rho(j)$ are absolutely summable, $T^{-1}|\widetilde{\Lambda}\widetilde{F}^r| \to 0$ for all $r = 1 - gT^{-1}$ with $g \ge 0$, $\mathrm{r}(\widetilde{V}) = O(1)$, $\mathrm{r}(\widetilde{V}^{-1}) = O(1)$, $T^{-2}\hat{u}'_{-1}(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\hat{u}_{-1} \overset{p}{\to} 0$, and $2T^{-1}\Delta\hat{u}'\widetilde{V}^{-1}\hat{u}_{-1} + 1 - \omega^{-2}[2T^{-1}\Delta\hat{u}'\hat{u}_{-1} + \gamma(0)] \overset{p}{\to} 0$.

We proceed by first proving that $T^{-1/2}|\widetilde{\Lambda}\widetilde{B}| \to 0$. With B_t being the tth row of $\widetilde{B}$, define the $T \times 3$ matrix $\widetilde{C}$ whose tth row is $C_t = T(B_{t+1} - B_t) = (0, 1, -\gamma\rho^t)$ and $C_T = 0$, so that $\widetilde{B} = T^{-1}\widetilde{F}^1\widetilde{C} + \tilde{e}B_1$. Now

$$\begin{aligned}
T^{-1}|\widetilde{\Lambda}\widetilde{B}|^2 &= T^{-1}\ \mathrm{tr}[\widetilde{B}'\widetilde{\Lambda}^2\widetilde{B}] \\
&= T^{-2}\ \mathrm{tr}[(\widetilde{B} + \tilde{e}B_1)'\widetilde{\Lambda}^2\widetilde{F}^1\widetilde{C}] + T^{-1}B_1B_1'\tilde{e}'\widetilde{\Lambda}^2\tilde{e} \\
&\le T^{-2}\omega|\mathrm{tr}[(\widetilde{B} + \tilde{e}B_1)'\widetilde{V}^{-1/2}\widetilde{\Lambda}\widetilde{F}^1\widetilde{C}]| + T^{-2}\omega^{-1}|\mathrm{tr}[(\widetilde{B} + \tilde{e}B_1)'\widetilde{V}^{1/2}\widetilde{\Lambda}\widetilde{F}^1\widetilde{C}]| \\
&\quad + T^{-1}B_1B_1'\tilde{e}'\widetilde{\Lambda}^2\tilde{e} \\
&\le T^{-2}\omega\ \mathrm{r}(\widetilde{V}^{-1/2})|\widetilde{\Lambda}\widetilde{F}^1||\widetilde{C}(\widetilde{B} + \tilde{e}B_1)'| + T^{-2}\omega^{-1}\ \mathrm{r}(\widetilde{V}^{1/2})|\widetilde{\Lambda}\widetilde{F}^1||\widetilde{C}(\widetilde{B} + \tilde{e}B_1)'| \\
&\quad + T^{-1}B_1B_1'|\tilde{e}'(\omega^{-2}\widetilde{V} + \omega^2\widetilde{V}^{-1} - 2\widetilde{I})\tilde{e}|.
\end{aligned}$$

But $|\widetilde{C}(\widetilde{B} + \tilde{e}B_1)'| = \ \mathrm{tr}^{1/2}|\widetilde{C}'\widetilde{C}(\widetilde{B} + \tilde{e}B_1)'(\widetilde{B} + \tilde{e}B_1)| = O(T)$ and $T^{-1}|\widetilde{\Lambda}\widetilde{F}^1| \to 0$ by the result of ERS, so that we are left to show that the final element in the sum above converges to zero, too.

With $\omega^{-2} = \sum_{j=-\infty}^{\infty}\rho(j)$ and $\omega^2 = \sum_{j=-\infty}^{\infty}\gamma(j)$ we find

$$\begin{aligned}
T^{-1}|\tilde{e}'(\widetilde{\Psi} - \omega^{-2}\widetilde{I})\tilde{e}| &= T^{-1}\left|2\sum_{j=1}^{T-1}(T-j)\rho(j) - 2T\sum_{j=1}^{\infty}\rho(j)\right| \\
&= 2\left|T^{-1}\sum_{j=1}^{T-1}j\rho(j) + \sum_{j=T}^{\infty}\rho(j)\right| \\
&\le 2\sum_{j=1}^{\infty}\min\left(\frac{j}{T}, 1\right)|\rho(j)| \to 0
\end{aligned}$$

and

$$T^{-1}|\tilde{e}'(\widetilde{V} - \omega^2\widetilde{I})\tilde{e}| \le 2\sum_{j=1}^{\infty}\min\left(\frac{j}{T}, 1\right)|\gamma(j)| \ \to 0$$

from the absolute summability of the sequences $\rho(j)$ and $\gamma(j)$. The result now follows from Lemma A1 of ERS.

(i) With the help of the result just established, we find

$$T^{-1}|\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\widetilde{B}| = T^{-1}\omega^{-1}|\widetilde{B}'\widetilde{V}^{-1/2}\widetilde{\Lambda}\widetilde{B}| \le T^{-1}\omega^{-1}|\widetilde{B}'|\mathrm{r}(\widetilde{V}^{-1/2})|\widetilde{\Lambda}\widetilde{B}| \to 0$$

since $\mathrm{tr}[\widetilde{B}'\widetilde{B}] = O(T)$.

(ii) We first treat the case $\gamma > 0$. With $\rho\tilde{u}_{-1} = \widetilde{F}^{\rho}\tilde{\nu} + \alpha\omega(2\gamma)^{-1/2}T^{1/2}\ \widetilde{B}(-\rho, 0, 1)'$, we find

$$\begin{aligned} T^{-3/2}\rho\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\tilde{u}_{-1} &= T^{-3/2}\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\widetilde{F}^{\rho}\tilde{\nu} \\ &\quad + T^{-1}\alpha(2\gamma)^{-1/2}\omega\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\widetilde{B}(-\rho, 0, 1)'. \end{aligned}$$

The second term goes to zero by (i). The trace of the covariance matrix of the first term is

$$\begin{aligned} \text{tr}[\text{var}[T^{-3/2}\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\widetilde{F}^{\rho}\tilde{\nu}]] &= T^{-3}\omega^{-2}\ \text{tr}[\widetilde{B}'\widetilde{V}^{-1/2}\widetilde{\Lambda}\widetilde{F}^{\rho}\widetilde{V}\widetilde{F}^{\rho\prime}\widetilde{\Lambda}\widetilde{V}^{-1/2}\widetilde{B}] \\ &\le T^{-3}\omega^{-2}|\widetilde{\Lambda}\widetilde{F}^{\rho}|^2|\widetilde{B}\widetilde{B}'|\ \text{r}^2(\widetilde{V}^{-1/2})\text{r}(\widetilde{V}) \to 0 \end{aligned} \tag{A.1}$$

since $T^{-1}|\widetilde{B}\widetilde{B}'| = O(1)$. For $\gamma = 0, \tilde{u}_{-1} = \widetilde{F}^1\tilde{\nu}$, so that the result follows from (A.1) with $\rho = 1$.

(iii) As

$$\text{tr}[\text{var}[T^{-1/2}\widetilde{B}'(\widetilde{V}^{-1} - \omega^{-2}\widetilde{I})\tilde{\nu}]] = T^{-1}\omega^{-2}\ \text{tr}[\widetilde{B}'\widetilde{\Lambda}^2\widetilde{B}] = T^{-1}\omega^{-2}|\widetilde{\Lambda}\widetilde{B}|^2 \to 0$$

and $\Delta\tilde{u} = \tilde{\nu} - \gamma T^{-1}\tilde{u}_{-1} - \alpha\omega 2^{-1/2}\gamma^{1/2}T^{-1/2}\tilde{e}$, the result follows using parts (i) and (ii).

(iv) and (v) For $\gamma = 0, \hat{u} = \tilde{u}$, and we are back to the analysis of ERS. For $\gamma > 0$, we have $\tilde{u}_{-1} = \hat{u}_{-1} + \alpha\omega(2\gamma)^{-1/2}T^{1/2}\widetilde{B}(-1, 0, \rho^{-1})'$ and $\Delta\tilde{u} = \Delta\hat{u} - \alpha\omega\gamma(2\gamma)^{-1/2}T^{-1/2}\rho^{-1}\widetilde{B}(0, 0, 1)'$, so the result follows applying parts (i), (ii), and (iii) to the respective pieces. *Q.E.D.*

LEMMA 2: *Let $\tilde{b}$ be a $T \times 1$ vector. If the elements b_t of $\tilde{b}$ satisfy* $\sup_t|b_t| = O_p(1)$, *then under Condition* 1 $\tilde{b}'\widetilde{V}^{-1}\tilde{\eta} = O_p(1)$. *Furthermore,* $\Delta\tilde{u}'\widetilde{V}^{-1}\tilde{\eta} = O_p(1)$.

PROOF: The proof will be carried out in the framework developed in the Appendix of ERS and already employed in Lemma 1. Define the $T \times T$ matrix $\widetilde{D} = \widetilde{I} - \widetilde{\Psi}\widetilde{V}$. For a real $T \times p$ matrix $K = [K_{tj}]$, let $\|K\| = \sum_{t=1}^{T}\sum_{j=1}^{p}|K_{tj}|$. Then $|K| \le \|K\|$. Furthermore, ERS argue at the beginning of their Appendix that under Condition 1, $\sum_{j=0}^{\infty}|j\gamma(j)| < \infty$ and $\|\widetilde{D}\| = O(1)$. Note that these inequalities imply that the sequence η_t is absolutely summable, since

$$\sum_{t=1}^{\infty}|\eta_t| = \sum_{t=1}^{\infty}\left|\sum_{j=0}^{\infty}r^j\gamma(j+t)\right| \le \sum_{t=1}^{\infty}\sum_{j=0}^{\infty}|\gamma(j+t)| = \sum_{j=0}^{\infty}|j\gamma(j)| < \infty.$$

Now

$$|\tilde{b}'\widetilde{V}^{-1}\tilde{\eta}| = |\tilde{b}'(\widetilde{\Psi} + \widetilde{D}\widetilde{V}^{-1})\tilde{\eta}| \le |\tilde{b}'\widetilde{\Psi}\tilde{\eta}| + |\tilde{\eta}'\widetilde{V}^{-1}\widetilde{D}'\tilde{b}|.$$

Since $\sup_t|b_t| = O_p(1)$ and the sequences $\rho(j)$ and η_t are absolutely summable, we have that $|\tilde{b}'\widetilde{\Psi}\tilde{\eta}| = O_p(1)$.

For the second term, first note that $\sup_t|b_t| = O_p(1)$ together with $\|\widetilde{D}\| = O(1)$ implies $\|\widetilde{D}'\tilde{b}\| = O_p(1)$. Furthermore, the absolute summability of the sequence η_t implies boundedness of $\tilde{\eta}'\tilde{\eta}$. We hence find

$$\begin{aligned} |\tilde{\eta}'\widetilde{V}^{-1}\widetilde{D}'\tilde{b}| &\le |\tilde{\eta}'\widetilde{V}^{-1}\|\widetilde{D}'\tilde{b}| \le |\tilde{\eta}'|\text{r}(\widetilde{V}^{-1})\|\widetilde{D}'\tilde{b}\| \\ &= (\tilde{\eta}'\tilde{\eta})^{1/2}\text{r}(\widetilde{V}^{-1})\|\widetilde{D}'\tilde{b}\| = O_p(1). \end{aligned}$$

For the second claim of the lemma, note that $\Delta\tilde{u} = \tilde{\nu} - \gamma T^{-1}\tilde{u}_{-1} - \alpha\omega 2^{-1/2}\gamma^{1/2}T^{-1/2}\tilde{e}$. But $(\gamma T^{-1}\tilde{u}_{-1} + \alpha\omega 2^{-1/2}\gamma^{1/2}T^{-1/2}\tilde{e})'\widetilde{V}^{-1}\tilde{\eta} = O_p(1)$ because $T^{-1}\tilde{u}_{-1}$ and $T^{-1/2}\tilde{e}$ satisfy the conditions for the vector $\tilde{b}$ of the lemma, and $\text{var}[\tilde{\nu}'\widetilde{V}^{-1}\tilde{\eta}] = \tilde{\eta}'\widetilde{V}^{-1}\tilde{\eta} = O_p(1)$ from another application of the first claim of the lemma, which concludes the proof. *Q.E.D.*

PROOF OF THEOREM 3: First note that since $G_iZ = 0$, $y'G_iy = u'G_iu$. The proof is for the time trend case $X = \tau$, the mean only case is a special simplified case using $Z = e$ and involves only the first element of each vector and matrix below.

(i) Let $r = 1 - cT^{-1}$ for $c = 0, g$. Define $\tilde{\varphi}(c) = \Delta\tilde{u} + cT^{-1}\tilde{u}_{-1}$ and $\widetilde{H}(c) = \tilde{e} + cT^{-1}\tilde{\tau}_{-1}$. Then $A(r)u = (0, \tilde{\varphi}(c)')'$, $A(r)e = (1, cT^{-1}\tilde{e}')'$, and $A(r)\tau = (0, \widetilde{H}(c)')'$. Since the first element of u

is zero, from direct calculation $u'(\Sigma_1^- - \Sigma_0^-)u = 2gT^{-1}\Delta\tilde{u}'\widetilde{V}^{-1}\tilde{u}_{-1} + g^2T^{-2}\tilde{u}'_{-1}\widetilde{V}^{-1}\tilde{u}_{-1}$. Noting that $2T^{-1}\Delta\tilde{u}'\tilde{u}_{-1} = T^{-1}u_T^2 - T^{-1}\sum_{t=1}^{T}\Delta u_t^2 \Rightarrow \omega^2 M(1)^2 - \gamma(0)$, an application of parts (iv) and (v) of Lemma 1, (3.3), and the CMT yield $u'(\Sigma_1^- - \Sigma_0^-)u \Rightarrow gM(1)^2 - g + g^2\int M(s)^2\,ds$. Furthermore, let $\Upsilon_k = \operatorname{diag}(k^{1/2}T^{1/2}, T^{-1/2})$ for $k > 0$, $\Upsilon_k = \operatorname{diag}(T^{1/2}, T^{-1/2})$ for $k = 0$, and $\overline{\Upsilon}_0 = \operatorname{diag}(1, T^{-1/2})$. Then

$$\Upsilon_k Z'\Sigma_1^- Z\Upsilon_k = \begin{pmatrix} 2g\omega^{-2} + g^2kT^{-1}\tilde{e}'\widetilde{V}^{-1}\tilde{e} + o(1) & gT^{-1}k^{1/2}\tilde{e}'\widetilde{V}^{-1}\widetilde{H}(g) \\ gT^{-1}k^{1/2}\widetilde{H}(g)'\widetilde{V}^{-1}\tilde{e} & T^{-1}\widetilde{H}(g)'\widetilde{V}^{-1}\widetilde{H}(g) \end{pmatrix}$$

and similarly $\overline{\Upsilon}_0 Z'\Sigma_0^- Z\overline{\Upsilon}_0 = \operatorname{diag}(1, T^{-1}\widetilde{H}(0)'\widetilde{V}^{-1}\widetilde{H}(0))$, $\Upsilon_k Z'\Sigma_1^- u = (gk^{1/2}T^{-1/2}\tilde{e}'\widetilde{V}^{-1}\tilde{\varphi}(g), T^{-1/2}\times \widetilde{H}(g)'\widetilde{V}^{-1}\tilde{\varphi}(g))'$, and $\overline{\Upsilon}_0 Z'\Sigma_0^- u = (0, T^{-1/2}\widetilde{H}(0)'\widetilde{V}^{-1}\ \tilde{\varphi}(0))'$. From Lemma 1 and direct calculations, we find $T^{-1}\tilde{e}'\widetilde{V}^{-1}\tilde{e} \to \omega^{-2}, T^{-1}\tilde{e}'\widetilde{V}^{-1}\widetilde{H}(c) \to \omega^{-2}\int(1+cs)ds, T^{-1}\widetilde{H}(c)'\widetilde{V}^{-1}\widetilde{H}(c) \to \omega^{-2}\int(1+cs)^2\,ds, T^{-1/2}\tilde{e}'\widetilde{V}^{-1}\tilde{\varphi}(c) - T^{-1/2}\omega^{-2}\tilde{e}'\tilde{\varphi}(c) \overset{p}{\to} 0$, and $T^{-1/2}\widetilde{H}(c)'\widetilde{V}^{-1}\tilde{\varphi}(c) - T^{-1/2}\omega^{-2}\widetilde{H}(c)'\tilde{\varphi}(c) \overset{p}{\to} 0$. Now

$$\begin{aligned} \widetilde{H}(c)'\tilde{\varphi}(c) &= \widetilde{H}(c)'\Delta\tilde{u} + cT^{-1}\widetilde{H}(c)'\tilde{u}_{-1} \\ &= (1+c)u_T + cT^{-1}(\widetilde{H}(c) - \tilde{e})'\tilde{u}_{-1} + o_p(1) \end{aligned}$$

and similarly $\tilde{e}'\tilde{\varphi}(c) = u_T + cT^{-1}\tilde{e}'\tilde{u}_{-1}$. The application of (3.3) and the CMT hence yields $T^{-1/2}\omega^{-1}\widetilde{H}(c)'\tilde{\varphi}(c) \Rightarrow (1+c)M(1) + c^2\int sM(s)\,ds$ and $T^{-1/2}\omega^{-1}\tilde{e}'\tilde{\varphi}(c) \Rightarrow M(1) + c\int M(s)ds$. Since $Q_a^\tau(g,k)$ is invariant to a time trend, we can substitute $M(s)$ with its projection off s, $M^\tau(s) = M(s) - 3s\int \lambda M(\lambda)\,d\lambda$. From the above results and the joint convergence of the separate pieces we find

$$\begin{aligned} Q_a^\tau(g,k) \Rightarrow{}& (g+1)M^\tau(1)^2 - g + g^2\int M^\tau(s)^2ds - \begin{pmatrix} k^{1/2}(g^2\int M^\tau(s)ds + gM^\tau(1)) \\ (g+1)M^\tau(1) \end{pmatrix}' \\ &\times \begin{pmatrix} 2g + g^2k & k^{1/2}(\frac{1}{2}g^2 + g) \\ k^{1/2}(\frac{1}{2}g^2 + g) & \frac{1}{3}g^2 + g + 1 \end{pmatrix}^{-1} \begin{pmatrix} k^{1/2}(g^2\int M^\tau(s)ds + gM^\tau(1)) \\ (g+1)M^\tau(1) \end{pmatrix}. \end{aligned}$$

The coefficients q_j^τ now follow after some algebra.

(ii) The only difference that arises between $\overline{Q}_a(g)$ and $Q_a(g,1)$ is through the terms $u'\Omega_1^{-1}u$, $Z'\Omega_1^{-1}Z$, and $Z'\Omega_1^{-1}u$. From the formula for partitioned inverses we find

$$\overline{V}^{-1} = \delta^{-1}\begin{pmatrix} 1 & -\tilde{\eta}'\widetilde{V}^{-1} \\ -\widetilde{V}^{-1}\tilde{\eta} & \delta\widetilde{V}^{-1} + \widetilde{V}^{-1}\tilde{\eta}\tilde{\eta}'\widetilde{V}^{-1} \end{pmatrix},$$

where $\delta = v_0(r) - \tilde{\eta}'\widetilde{V}^{-1}\tilde{\eta} = \omega^2T(2g)^{-1} + o(T)$ from Lemma 2. With $r = 1 - gT^{-1}$, $A(r)u = (0, \Delta\tilde{u}' + gT^{-1}\tilde{u}'_{-1})'$, $A(r)e = (1, gT^{-1}\tilde{e}')'$, and $A(r)\tau = (0, \tilde{e}' + gT^{-1}\tilde{\tau}'_{-1})'$; the additional terms of $\overline{Q}_a(g)$ compared to $Q_a(g,1)$ are of the form $T^{-1}\tilde{\eta}'\widetilde{V}^{-1}\tilde{\eta}, \delta^{-1}\tilde{\eta}'\widetilde{V}^{-1}\tilde{b}$, and $\delta^{-1}\tilde{b}'\widetilde{V}^{-1}\tilde{\eta}\tilde{\eta}'\widetilde{V}^{-1}\tilde{d}$, where $\tilde{b}$ and $\tilde{d}$ stand for either of $T^{-1/2}\tilde{u}_{-1}, \Delta\tilde{u}, \tilde{e}$ or $T^{-1}\tilde{\tau}_{-1}$. But $\sup_t|T^{-1/2}u_t| \Rightarrow \sup_s|\omega M(s)| = O_p(1)$, so that an application of Lemma 2 shows that all of these terms converge to zero in probability. *Q.E.D.*

METHOD TO FIND A COMPARABLE $Q_a(g,k)$ FOR A GIVEN CLASS OF TESTS: Denote with r_n the decision of a given class of tests of size 5% to reject ($r_n = 1$) or not to reject ($r_n = 0$) the nth draw $W_n^i(s)$ of a random sample $n = 1, \ldots, N$ of a detrended Wiener process. Consider a nonlinear logit regression of r_n on a constant and a scaled weighted sum of $W_n^i(1)^2$, $(\int W_n^i(s)ds)^2$, $W_n^i(1)\int W_n^i(s)ds$, and $\int W_n^i(s)^2ds$, where the weights depend on g and k and are given by the weights q_j^i of $Q_a^i(g,k)$ (cf. Theorem 3),

$$r_n = L\left(l_0 + l_1\left[q_1^iW_n^i(1)^2 + q_2^i\int W_n^i(s)ds^2 + q_3^iW_n^i(1)\int W_n^i(s)ds + q_4^i\int W_n^i(s)^2ds\right]\right) + e_n,$$

where $L(x)$ is the logistic function $L(x) = 1/(1+e^{-x})$, and the estimated parameters are l_0, l_1, g, and k. Then the estimated values of g and k in this regression may serve as approximations to the values of g and k that maximize the asymptotic probability that the two tests both reject or do not reject under the null hypothesis of $\rho = 1$. The values of Table II were calculated with $N = 80{,}000$.

REFERENCES

ANDREWS, D., AND W. PLOBERGER (1994): "Optimal Tests When a Nuisance Parameter is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

BHARGAVA, A. (1986): "On the Theory of Testing for Unit Roots in Observed Time Series," *Review of Economic Studies*, 53, 369–384.

CHAN, N., AND C. WEI (1987): "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *The Annals of Statistics*, 15, 1050–1063.

DICKEY, D., AND W. FULLER (1979): "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, 74, 427–431.

DUFOUR, J.-M., AND M. KING (1991): "Optimal Invariant Tests for the Autocorrelation Coefficient in Linear Regressions with Stationary or Nonstationary AR(1) Errors," *Journal of Econometrics*, 47, 115–143.

ELLIOTT, G. (1999): "Efficient Tests for a Unit Root When the Initial Observation is Drawn From its Unconditional Distribution," *International Economic Review*, 40, 767–783.

ELLIOTT, G., T. ROTHENBERG, AND J. STOCK (1996): "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64, 813–836.

EVANS, G., AND N. SAVIN (1981): "Testing for Unit Roots: 1," *Econometrica*, 49, 753–779.

——— (1984): "Testing for Unit Roots: 2," *Econometrica*, 52, 1241–1269.

LEHMANN, E. (1986): *Testing Statistical Hypotheses*, Second Edn. New York: Wiley.

MÜLLER, U. (2002): "Tests for Unit Roots and the Initial Observation," Ph.D. Thesis, University of St. Gallen.

NABEYA, S., AND K. TANAKA (1990): "Limiting Power of Unit-Root Tests in Time-Series Regression," *Journal of Econometrics*, 46, 247–271.

PANTULA, S., G. GONZALEZ-FARIAS, AND W. FULLER (1994): "A Comparison of Unit-Root Test Criteria," *Journal of Business & Economic Statistics*, 12, 449–459.

PHILLIPS, P. (1987a): "Time Series Regression with a Unit Root," *Econometrica*, 55, 277–301.

——— (1987b): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547.

PHILLIPS, P., AND P. PERRON (1988): "Testing for a Unit Root in Time Series Regression," *Biometrika*, 75, 335–346.

RAO, C., AND S. MITRA (1971): *Generalized Inverse of Matrices and its Applications*. New York: Wiley.

ROTHENBERG, T., AND J. STOCK (1997): "Inference in a Nearly Integrated Autoregressive Model with Nonnormal Innovations," *Journal of Econometrics*, 80, 269–286.

SARGAN, J., AND A. BHARGAVA (1983): "Testing Residuals from Least Squares Regression for Being Generated by the Gaussian Random Walk," *Econometrica*, 51, 153–174.

SCHMIDT, P., AND J. LEE (1991): "A Modification of the Schmidt-Phillips Unit Root Test," *Economics Letters*, 36, 285–289.

SCHMIDT, P., AND P. PHILLIPS (1992): "LM Tests for a Unit Root in the Presence of Deterministic Trends," *Oxford Bulletin of Economics and Statistics*, 54, 257–287.

STOCK, J. (1994): "Unit Roots, Structural Breaks and Trends," in *Handbook of Econometrics*, Vol. 4, ed. by R. Engle and D. McFadden. New York: North Holland, pp. 2740–2841.

——— (2000): "A Class of Tests for Integration and Cointegration," in *Cointegration, Causality, and Forecasting—A Festschrift in Honour of Clive W. J. Granger*, ed. by R. Engle and H. White. Oxford: Oxford University Press, pp. 135–167.