



HIER

Harvard Institute of Economic Research

Discussion Paper Number 2003

Optimal Versus Robust Inference in Nearly Integrated
Non Gaussian Models

by

Samuel B. Thompson

April 2003

Harvard University
Cambridge, Massachusetts

This paper can be downloaded without charge from:

<http://post.economics.harvard.edu/hier/2003papers/2003list.html>

Optimal versus robust inference in nearly integrated non Gaussian models

Samuel B. Thompson
Department of Economics
Harvard University
Cambridge, MA 02138

January 10, 2003

Abstract: Elliott, Rothenberg and Stock (1996) derived a class of point-optimal unit root tests in a time series model with Gaussian errors. Other authors have proposed “robust” tests which are not optimal for any model but perform well when the error distribution has thick tails. I derive a class of point-optimal tests for models with non Gaussian errors. When the true error distribution is known and has thick tails, the point-optimal tests are generally more powerful than Elliott et al.’s (1996) tests as well as the robust tests. However, when the true error distribution is unknown and asymmetric, the point-optimal tests can behave very badly. Thus there is a trade-off between robustness to unknown error distributions and optimality with respect to the trend coefficients.

Keywords: Near unit root, robust tests, optimal tests

This paper could not have been written without the encouragement of Thomas Rothenberg. This is based on my dissertation, which he supervised. I also thank Don Andrews, Jack Porter, Jim Stock, and seminar participants at the University of Pennsylvania, the University of Toronto, the University of Montreal, Princeton University and the meetings of the Econometric Society in UCLA. Comments of three anonymous referees greatly improved the exposition of the paper. I owe special thanks to Gary Chamberlain for helping me to understand these results.

1 Introduction

Elliott et al. (1996) derived a class of point-optimal unit root tests in a time series model with Gaussian errors. They showed that, by efficiently handling intercept and trend coefficients, their tests are generally more powerful than the standard Dickey-Fuller tests. The present paper investigates whether the same power improvements can be attained when using “robust” testing methods that are designed to improve power for non Gaussian error distributions. I find that this improvement occurs when the true error distribution is known or at least is known to be symmetric. However, if one wants to be robust to thick-tailed, possibly asymmetric error distributions, the power improvement found by Elliott et al. (1996) cannot be attained.

First I consider the model with an intercept and no time trend. In large samples the variation of a nearly integrated process dominates the intercept of the process. Thus the intercept can be set equal to zero when forming test statistics. The resulting point-optimal tests dominate previously proposed robust tests (see Lucas (1995), Hecce (1996), and Hasan and Koenker (1997)) which do not set the intercept to zero.

However, when the error distribution is unknown and asymmetric, setting the intercept to zero leads to a test with very bad properties. In large samples the zero-intercept tests reject a true null hypothesis with probability approaching one-half. The previously proposed inefficient tests perform well under asymmetric errors. Thus there is a trade-off between efficiently handling conditioning variables and robustness with respect to asymmetric error distributions.

Then I consider the model with both an intercept and a linear time trend. I derive the form of the point-optimal test which is invariant to the time trend. In many cases it is difficult to compute the point-optimal test, so I use Laplace’s approximation to derive an asymptotically equivalent test which is easier to calculate. I show that tests based on the maximum likelihood estimator (mle) and the likelihood ratio (lr) statistic, which were previously studied by Xiao (2001), are asymptotically admissible. When the error distribution is known and non Gaussian, a test based on either of these statistics will in many cases have higher power than the tests suggested by Elliott et al. (1996).

In the model with a time trend, an unknown asymmetric error distribution causes the power of the point-optimal test to approach zero in large samples. The tests based on the mle and lr statistic have slightly better properties - they have power approaching zero against local alternatives, but power approaching 1 against fixed alternatives. Thus, while asymmetric errors lead to power losses for these two procedures, the tests do not over reject a true null and are acceptable for both correct and incorrectly specified errors. Monte Carlo results suggest the power losses are substantial for the point-optimal tests, but not as bad for the the mle and lr tests.

Thus the viable unit root tests are the traditional robust tests (which inefficiently handle intercepts and trends) and the point-optimal Gaussian tests proposed in Elliott et al. (1996) (which are inefficient in the presence of thick-tailed errors). In some situations the efficiency loss due to ignoring thick-tailed errors is less than that due to inefficiently modelling the intercept and trend. For example, the point-optimal Gaussian test is more powerful than many traditional robust tests when the errors

are drawn from a Student's t distribution with five or more degrees of freedom.

Although the present paper does not specifically consider the topic, there are similar implications for the construction of confidence intervals for autoregressive roots close to one.¹ Since many of the intervals are based on the inversion of tests, it appears that the framework for constructing more accurate intervals described in Elliott and Stock (2001) cannot be extended to non Gaussian models.

2 The model with no time trend

The observations $\{y_t\}_{t=1}^T$ come from the model

$$\begin{aligned} y_t &= \beta' x_t + u_t \\ \Delta u_t &= \gamma u_{t-1} + \varepsilon_t \end{aligned}$$

where $\beta = (\beta_1, \beta_2)'$ is a 2-dimensional coefficient vector and $x_t = (1, t)'$. I consider the model with an intercept only (e.g., $\beta_2 = 0$), and with a linear time trend (e.g., no restrictions on β). The random errors ε_t are iid and have expectation zero and a finite variance. Under the unit root hypothesis, $\gamma = 0$ and the detrended series is not stationary. I will evaluate tests of the unit root hypothesis versus the alternatives $\gamma < 0$. Since I am interested in inference when γ is close to one, I adopt the local-to-zero reparameterization $\gamma = c/T$, so the parameter space is a shrinking neighborhood of zero as the sample size grows. Following Chan and Wei (1987) and Phillips (1987), I take c fixed when making limiting arguments, obtaining asymptotic power as a function of the local alternative c .

We distinguish between the true, unknown density for ε , given by $e^{-f(\varepsilon)}$, and the density used to construct the likelihood function, $e^{-g(\varepsilon)}$. The researcher chooses g hoping that g is a reasonable approximation to f , and also hoping that the resulting tests perform well when $g \neq f$. In the model with an intercept and no time trend,

$$L(c, \beta_1) = \sum_{t=2}^T g \left(\Delta y_t - \frac{c}{T} y_{t-1} + \beta_1 \frac{c}{T} \right) \quad (1)$$

is the negative of the log-likelihood function evaluated at $\gamma = c/T$, conditional on the first observation y_1 .

Consider the classical regression model $y = \alpha_0 + \alpha_1 x + \varepsilon$ with nonrandom x and iid error ε . If the true value of the intercept α_0 is zero, then regressing y on x alone leads to a more efficient estimator of α_1 than regressing y on both x and a constant. Now consider two estimators for c :

1. $(\tilde{c}, \tilde{a}) = \operatorname{argmin}_{(c,a)} \sum g(\Delta y_t - cy_{t-1}/T - a)$, with $a = -\beta_1 c/T$. These are the usual M-estimators studied by Lucas (1995), Hoek, Lucas and van Dijk (1995),

¹Methods for constructing these intervals appear in Stock (1991), Hansen (1999), and Elliott and Stock (2001).

Herce (1996), and Hasan and Koenker (1997).²

2. $\hat{c} = \operatorname{argmin}_c \sum g(\Delta y_t - cy_{t-1}/T)$. I label this statistic the “constrained” mle.

If β_1 is zero then a is zero and \hat{c} should be more efficient than \tilde{c} . Thus a test which rejects the null for small values of \hat{c} should be more powerful than a test which rejects for small \tilde{c} .

We include the constant a in case β_1 is not zero. However in large samples $a = -\beta_1 c/T$ is very close to zero no matter what the true values for β_1 and c . This suggests that asymptotically it does not matter that we omit the constant. It turns out that, if g equals f (the true negative log-density of the errors), then in large samples tests based on \hat{c} dominate tests based on \tilde{c} even when β_1 and c are not zero.

This is the source of the power improvements in the model with no time trend. Many existing robust unit root tests do not take advantage of the fact that in large samples the variation in u_t dominates any fixed intercept, so β_1 can be taken equal to zero without affecting the asymptotic distribution of \hat{c} .

We will show that in large samples, no test dominates the test based on \hat{c} . This optimality result comes from the Neyman-Pearson lemma, which states that the most powerful test of $c = 0$ versus the alternative $c = \bar{c}$ rejects for small values of $L(\bar{c}, \beta_1) - L(0, \beta_1)$. In large samples, the \hat{c} -test is just as powerful as the Neyman-Pearson statistic for some \bar{c} . This is true even when β_1 is not known.

Elliott et al. (1996) have shown that in a Gaussian model with an intercept and no time trend, there is no efficiency loss from β_1 being unknown. The same is true for nonnormal innovations. Suppose we form the Neyman-Pearson test with an incorrect value for β_1 , say 0. If g is three times differentiable with bounded second and third derivatives then by a Taylor series approximation,

$$\begin{aligned} L(\bar{c}, 0) - L(0, 0) &= \sum g\left(\varepsilon_t - \frac{\bar{c} - c}{T}u_{t-1} + \frac{\bar{c}}{T}\beta_1\right) - \sum g\left(\varepsilon_t + \frac{c}{T}u_{t-1}\right) \\ &= -\frac{\bar{c}}{T} \sum g'(\varepsilon_t) u_{t-1} + \frac{\bar{c}^2 - 2\bar{c}c}{2T^2} \sum g''(\varepsilon_t) u_{t-1}^2 \\ &\quad + \bar{c}\beta_1 \left[\frac{1}{T} \sum g'(\varepsilon_t) + \frac{1}{T^2} \sum g''(\varepsilon_t) \left(\frac{\bar{c}\beta_1}{2} - (\bar{c} - c)u_{t-1} \right) \right] \\ &\quad - \frac{1}{6T^3} \sum g'''(\varepsilon_t^*) ((\bar{c} - c)u_{t-1} - \bar{c}\beta_1)^3 - \frac{1}{6T^3} \sum g'''(\varepsilon_t^{**}) (cu_{t-1})^3, \end{aligned}$$

where $|\varepsilon_t^* - \varepsilon_t| \leq |(\bar{c} - c)u_{t-1}/T - \bar{c}\beta_1/T|$ and $|\varepsilon_t^{**} - \varepsilon_t| \leq |cu_{t-1}/T|$. Under regularity conditions given below, $u_{t-1}/T^{1/2}$ is $O_p(1)$. Therefore, since g'' and g''' are bounded, many of the terms are asymptotically negligible:

$$\underline{L(\bar{c}, 0) - L(0, 0) = -\frac{\bar{c}}{T} \sum g'(\varepsilon_t) u_{t-1} + \frac{\bar{c}^2 - 2\bar{c}c}{2T^2} \sum g''(\varepsilon_t) u_{t-1}^2 + \frac{\bar{c}\beta_1}{T} \sum g'(\varepsilon_t) + o_p(1).}$$

²Hasan and Koenker (1997) proposed rank tests instead of M-tests. Thompson (2001b) notes that under the local-to-zero reparameterization, for each rank test and error distribution there exists a test based on \tilde{c} with the same asymptotic power function. Thus we will not specifically discuss the rank tests.

If $E g'(\varepsilon_t) = 0$, then $T^{-1} \sum g'(\varepsilon_t) \xrightarrow{p} 0$ and in large samples the test statistic does not depend on β_1 . $E g'(\varepsilon_t)$ will equal zero when the errors are correctly specified, meaning that $g = f$:

$$E g'(\varepsilon_t) = \int_{-\infty}^{\infty} f'(z) e^{-f(z)} dz = - \left. \frac{\partial}{\partial x} \int_{-\infty}^{\infty} e^{-f(z+x)} dz \right|_{x=0} = - \left. \frac{\partial}{\partial x} (1) \right|_{x=0} = 0. \quad (2)$$

Thus, under correct specification of the errors, there is no efficiency loss from β_1 being unknown.

In a stationary autoregressive model, the Neyman-Pearson test statistic typically admits an asymptotic representation in terms of a single scalar sufficient statistic. This allows the construction of a test which is asymptotically uniformly most powerful against all alternatives $c < 0$. Here the Neyman-Pearson statistic has an asymptotic representation that is a linear combination of the two scalar sufficient statistics $T^{-1} \sum g'(\varepsilon_t) u_{t-1}$ and $T^{-2} \sum g''(\varepsilon_t) u_{t-1}^2$, with weights depending on \bar{c} . As Elliott et al. (1996) have noted, this implies that there does not exist a uniformly most powerful test, even in large samples. Each Neyman-Pearson test is most powerful only against the point alternative $c = \bar{c}$. The Neyman-Pearson tests comprise an infinite family of admissible tests, indexed by \bar{c} , no one dominating the others for all c .

Since there is no uniformly most powerful test, the goal is to find feasible, admissible tests. Let $\pi(c, \bar{c})$ denote the asymptotic power function for the Neyman-Pearson test indexed by \bar{c} when the true value of the local autoregressive parameter is c and the size of the test is α :

$$\pi(c, \bar{c}) = \lim_{T \rightarrow \infty} \Pr [L(\bar{c}, \beta_1) - L(0, \beta_1) < q(\bar{c})]$$

where $q(\bar{c})$ satisfies $\pi(0, \bar{c}) = \alpha$. Since the Neyman-Pearson test indexed by \bar{c} is asymptotically optimal against the alternative $c = \bar{c}$, the envelope power function $\Pi(c) \equiv \pi(c, c)$ is the upper bound on power for all tests against each alternative. A test is asymptotically admissible if it has a limiting power function which is equal and tangent to the envelope function for some c .

In the next section I show that the \hat{c} test is asymptotically admissible while the \tilde{c} test is not. There are other interesting test statistics to consider: the M-estimator t -test, which rejects for small values of $\tilde{t} = [T^{-2} \sum (y_{t-1} - \bar{y})^2]^{1/2} \tilde{c}$, and the constrained t and likelihood ratio statistics

$$\hat{t} = \sqrt{T^{-2} \sum y_{t-1}^2} \hat{c} \quad \text{and} \quad \hat{l} = -2 \left[\min_c L(c, 0) - \min L(0, 0) \right].$$

The \hat{t} and \hat{l} tests impose the constraint $\beta_1 = 0$, so they will dominate the M-estimator t -test.

2.1 Asymptotic power functions

In order to justify the claim that the \hat{c} , \hat{t} and \hat{l} tests dominate the M-tests, it will prove convenient to develop asymptotic representations for the various statistics. Consider

some of the g functions used for robust regression problems:

Least Squares	$g(x) = x^2/2$
LAD regression	$g(x) = x $
q^{th} quantile regression	$g(x) = qx - x1(x < 0)$
Huber's M function	$g(x) = (x^2/2)1(x < k) + (k x - k^2/2)1(x \geq k)$

where the constant k is chosen by the researcher. Since g may not be everywhere differentiable, we cannot approximate the log-likelihood function with Taylor series expansions. Instead of pointwise differentiability, the proofs make use of “stochastic differentiability,” an idea described in Pollard (1985). Application of the idea requires that I impose smoothness conditions on the error density to make up for the lack of smoothness in the objective function.

Assumption 1 (*Smoothness of the error density*): The errors $\{\varepsilon_t\}_{t=1}^T$ are iid mean zero with $E|\varepsilon_1|^{2+\delta} < H$ for some $\delta > 0$. ε_1 has a density function $f(z)$ which is bounded and uniformly continuous.

The g function may have finitely many points of nondifferentiability.

Assumption 2 (*Objective function*): $g(x)$ is convex and strictly increasing in $|x|$. $g(x)$ is everywhere twice differentiable except for x in P , where P contains the D points p_1, \dots, p_D . There exists some finite positive H so that $|g''(x)| < H$ for x not in P . There exists some finite positive h satisfying $P \in [-h - \delta, h + \delta]$ for some $\delta > 0$, so that for all x and y in $[-h, h]$ we have $|g'(x)| < H$ and $|g(x) - g(y)| \leq H|x - y|$.

I assume that g is convex because it simplifies the proofs. Assuming convexity allows me to extend several pointwise convergence results to apply uniformly over the parameter space. Convexity also greatly simplifies the demonstration of the rate of convergence of the estimators. My extensive use of convexity is due to results in Pollard (1991) and Hjort and Pollard (1993).

For non differentiable g , it is not possible to define an approximate likelihood in terms of the derivatives g' and g'' . We replace g' with the derivative-like function ψ :

Definition 1 $\psi(x)$ is equal to $g'(x)$ if g is differentiable at x and $\psi(x) = 0$ otherwise.

If g is everywhere differentiable then $g' = \psi$. For LAD regression $\psi(x) = \text{sign}(x)$, and for Huber's function $\psi(x) = x1(|x| < k) + k1(|x| \geq k)$.

In standard (non unit-root) problems, the second derivative g'' enters the asymptotic representation though its expectation $Eg''(\varepsilon_1)$. We replace $Eg''(\varepsilon_1)$ with the parameter $\omega = -\int_{\mathbb{R}} \psi(x) df(x)$. When g is everywhere twice differentiable, $\omega = Eg''(\varepsilon_1)$. For LAD regression $\omega = 2f(0)$ and for Huber's function $\omega = \Pr[|x| < k]$.

In large samples the power functions admit representations in terms of functionals of Brownian motion. Define $W(\cdot)$ to be standard Brownian motion and define $W_c(\cdot)$ to be the Ornstein-Uhlenbeck process $W_c(t) = \int_0^t \exp\{c(t-s)\} dW(s)$ with initial condition $W_0(0) = 0$. The asymptotic representations make use of the parameters

$\sigma_\varepsilon^2 = \text{Var}(\varepsilon_t)$, $\rho = \text{Corr}(\varepsilon_t, \psi(\varepsilon_t))$ and $\sigma_\psi^2 = \text{Var}\psi(\varepsilon_t)$, as well as the stochastic process

$$S_\rho(t) = \rho W(t) + (1 - \rho)^{1/2} \widetilde{W}(t),$$

where \widetilde{W} is standard Brownian motion, independent of W . The following theorem is proven in the Appendix.

Theorem 1 *If $E\psi(\varepsilon_t) = 0$, and if assumptions (1) and (2) hold, then*

1. $L(\bar{c}, 0) - L(0, 0) \Rightarrow -\bar{c}\sigma_\varepsilon\sigma_\psi \int W_c dS_\rho + 2^{-1}(\bar{c}^2 - 2\bar{c}c)\omega\sigma_\varepsilon^2 \int W_c^2$,
2. $\hat{c} \Rightarrow \sigma_\psi [\omega\sigma_\varepsilon \int W_c^2]^{-1} \int W_c dS_\rho + c$,
3. $\hat{t} \Rightarrow (\sigma_\varepsilon^2 \int W_c^2)^{-1/2} (\sigma_\varepsilon\sigma_\psi \int W_c dS_\rho + \omega\sigma_\varepsilon^2 \int W_c^2)$,
4. $\hat{l} \Rightarrow \omega^{-1} (\sigma_\varepsilon^2 \int W_c^2)^{-1} (\sigma_\varepsilon\sigma_\psi \int W_c dS_\rho + \omega\sigma_\varepsilon^2 \int W_c^2)^2$,
5. $\tilde{c} \Rightarrow \sigma_\psi [\omega\sigma_\varepsilon \int D_c^2]^{-1} \int D_c dS_\rho + c$, where $D_c(r) = W_c(r) - \int W_c(s) ds$,
6. $\tilde{t} \Rightarrow (\sigma_\psi/\omega) [\int D_c^2]^{-1/2} \int D_c dS_\rho + c\sigma_\varepsilon [\int D_c^2]^{1/2}$.

Rothenberg and Stock (1997) and Xiao (2001) derived similar representations without assuming convexity of g , but did not allow for nondifferentiable functions.

The large sample power function for the Neyman-Pearson test is

$$\pi(c, \bar{c}) = \Pr \left[-\bar{c}\sigma_\psi\sigma_\varepsilon \int W_c dS_\rho + (\bar{c}^2 - 2\bar{c}c) \frac{\omega\sigma_\varepsilon^2}{2} \int W_c^2 < q(\bar{c}) \right].$$

Power functions for the other tests may be obtained similarly.

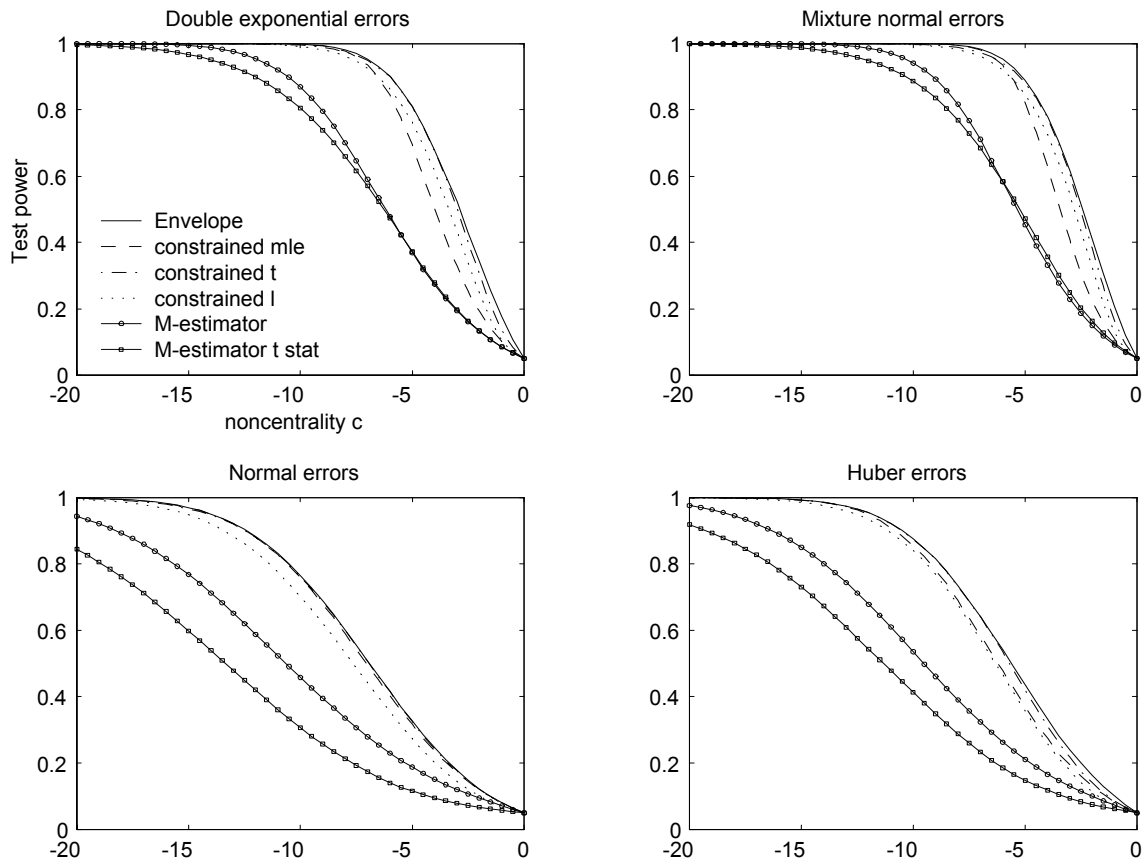
Figure 1 plots envelope power functions and asymptotic power for a variety of tests. The curves for LAD errors (from the double exponential distribution) are given, as well as standard normal errors and Huber errors³. I also produce a curve for the mixture distribution (labeled Mixture in the figure) where a standard normal variable is drawn with probability .95 and a $N(0, 36)$ variable is drawn with probability .05.⁴ Each curve is calculated under the assumption of correct specification, so that e^{-g} is equal to the true density e^{-f} .

The power curves for nonnormal errors are all substantially higher than the curve for normal errors. The most powerful test for Gaussian errors achieves 50% power at c close to -7.0 and the most powerful test for double exponential errors (corresponding to LAD estimation under correct specification) achieves 50% power at c close to -3.75 .

³The parameter k which appears in Huber's M function is set to 1.345 for all of the figures in this paper. At this value of k , the Huber estimate of a location parameter from iid standard normal data has a relative efficiency of 95% with respect to the mean. See Hampel, Ronchetti, Rousseeuw and Stahel (1986) page 399.

⁴The log of the density for the mixture distribution is not convex. While this violates the assumptions used to derive the asymptotic representations in the Appendix, simulations not reported here suggest that the representations are still valid.

Figure 1: Asymptotic power curves for unit root tests in the model with no time trend ($x_t = (1, 0)$). The curves are drawn under the assumption of correct specification, so the g function used to form the test statistics is equal to f , the negative log-density of the errors.



Notes: The simulations which appear in this paper were performed by computing stochastic integrals as the realizations of normalized sums of 500 successive draws from a discrete time Gaussian AR(1) process with autoregressive parameter $1 - c/T$. There are 100,000 Monte Carlo replications.

Figure also 1 provides power curves for the tests based on the constrained mle \hat{c} and the M-estimator \tilde{c} . The \hat{c} -test is asymptotically admissible. Test power is tangent to the power envelope when envelope power is large. The \tilde{c} -test is not asymptotically admissible and is dominated by the \hat{c} -test. The power curve for \tilde{c} touches the envelope function only under the null ($c = 0$) and for alternatives so far from zero that any sensible test would have power equal to 1.

The M-estimator t -test is not asymptotically admissible, while the constrained \hat{t} test is admissible. Figure 1 shows that the constrained t test achieves tangency to the power envelope function at power close to 50%. The Figure also shows that the constrained \hat{l} test is not admissible. As Rothenberg and Stock (1997) have shown, straightforward manipulations of the asymptotic representations demonstrate that rejecting for large values of \hat{l} is asymptotically equivalent to rejecting for large \hat{t}^2 .

Since the tests based on \hat{t} and \hat{l} are one-sided and two-sided tests of the same one-sided hypothesis, it is not surprising that the t-test dominates the test based on \hat{l} .

We have obtained power improvements by imposing the constraint that the intercept estimate $\hat{\beta}_1$ is zero. We can obtain identical results by replacing the requirement that $\hat{\beta}_1 = 0$ with the requirement that the estimator $\hat{\beta}_1$ is stochastically bounded. Consider two more estimators.

1. $(\hat{c}_{bound}, \hat{\beta}_{bound}) = \operatorname{argmin}_{(c, \beta_1 \in \Theta)} \sum_{t \geq 2} g(\Delta y_t - c y_{t-1}/T + \beta_1 c/T)$, where Θ is a compact set. It is common to assume a bounded parameter space, and this estimator imposes that assumption. The estimator $\hat{\beta}_{bound}$ is obviously stochastically bounded and the term $\sup_{\beta_1 \in \Theta} |\beta_1 c/T|$ disappears from the likelihood function in large samples. Thus \hat{c}_{bound} has the same limiting distribution as \hat{c} .
2. $(\hat{c}_{initial}, \hat{\beta}_{initial}) = \operatorname{argmin}_{(c, \beta_1)} \{g(y_1 - \beta_1) + \sum_{t \geq 2} g(\Delta y_t - c y_{t-1}/T + \beta_1 c/T)\}$. These are the maximum likelihood estimators when we assume the initial condition $u_0 = 0$. In an earlier draft of this paper I showed that $\hat{\beta}_{initial}$ is stochastically bounded. Thus $(\hat{c}_{initial}, \hat{\beta}_{initial})/T \rightarrow 0$ fast enough so that $\hat{c}_{initial}$ has the same distribution as \hat{c} .

2.2 Failure of robustness to error misspecification

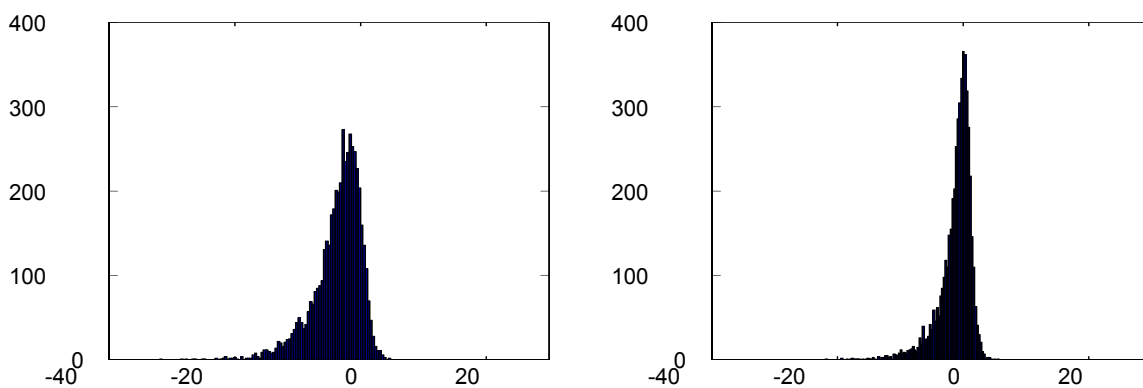
These power improvements occur so long as g , the estimating function, is equal to f , the true negative log-density of the errors. When $g \neq f$, \hat{c} may behave poorly. Consider the classical regression model $y = \alpha_0 + \alpha_1 x + \varepsilon$. If the errors come from the double exponential distribution, ε has zero median and the maximum likelihood estimates are the LAD estimates $(\tilde{\alpha}_0, \tilde{\alpha}_1) = \operatorname{argmin}_{(\alpha_0, \alpha_1)} \sum |y - \alpha_0 - \alpha_1 x|$. If the true value of α_0 is zero then under correct specification we can get a better estimator for α_1 by removing α_0 from the objective function: $\hat{\alpha}_1 = \operatorname{argmin}_{\alpha_1} \sum |y - \alpha_1 x|$. Now suppose that $\alpha_0 = 0$ and that ε comes from an incorrectly specified, asymmetric error distribution with zero mean but nonzero median. For example take $\varepsilon = Z^2 - 1$ where Z is standard normal. It is well known that in this case $\tilde{\alpha}_0 \xrightarrow{p} \operatorname{median}(\varepsilon)$ and $\tilde{\alpha}_1$ has a limiting distribution. It is also well known that the distribution of $\hat{\alpha}_1$ is not stochastically bounded, even if $\alpha_0 = 0$. Thus the constant α_0 ‘‘recenters’’ the incorrectly specified errors. In the classical setting we include the constant to protect ourselves from errors with nonzero median.

The same thing happens in the unit root problem. Hecce (1996) has shown that if the intercept β_1 is zero and the errors have zero median, then \tilde{c} and \hat{c} both have limiting distributions. When ε_t has a nonzero median \tilde{c} has a limiting distribution while \hat{c} blows up. This can be seen in the histograms in figure 2. Each histogram depicts 5000 Monte Carlo realizations of $\tilde{\alpha}_1$ and $\hat{\alpha}_1$, estimated from simulated data sets of 500 observations from the model with $c = 0$ (so the null is true), $\beta_1 = 0$ and initial condition $u_0 = 0$. The top two histograms show that when the errors come from the zero median Student’s t distribution with four degrees of freedom, both estimators have limiting distributions. The bottom two histograms show that when the errors do not have zero medians ($\varepsilon_t = Z_t^2 - 1$ where Z_t are iid standard normal)

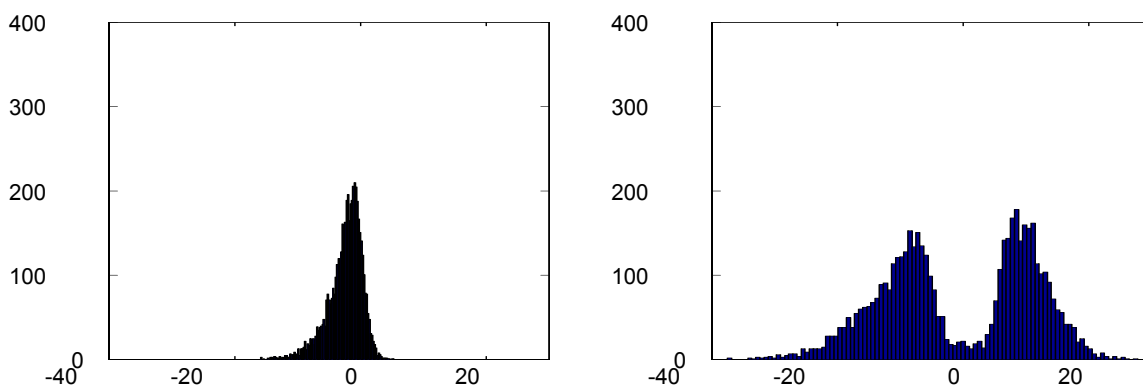
the distribution of \hat{c} blows up. So for any fixed critical value q , the probability of rejecting a true null hypothesis converges to $\lim_{T \rightarrow \infty} \Pr[\hat{c} < q] = .5$.

Figure 2: Histograms of 5000 Monte Carlo simulations of \tilde{c} (on the left) and \hat{c} (on the right) estimated from 500 observations from the model with no trend or intercept. The true value of c is 0.

Here the errors come from the Student's t4 distribution. Since the median is zero \hat{c} is more efficient and has a smaller spread around 0.



Here each error is $\varepsilon_t = Z_t^2 - 1$, where Z_t are iid standard normal. Since the median is not zero, \tilde{c} has a limiting distribution and \hat{c} blows up.



Mathematically this can be understood as failure of an identification condition. Consider the classical regression model with no intercept: $y = \alpha_1 x + \varepsilon$. If g is differentiable, the estimator $\hat{\alpha}_1 = \operatorname{argmin}_{\alpha_1} \sum g(y - \alpha_1 x)$ will solve the first order condition $T^{-1} \sum g'(y - \hat{\alpha}_1 x) x = 0$. Under the usual assumptions this will deliver a consistent estimator of α_1 if the identification condition $E g'(y - \alpha_1 x) x = 0$ holds. Since $y - \alpha_1 x = \varepsilon$ and x is not random, this condition is equivalent to $E g'(\varepsilon) = 0$. Thus the identification condition is that ε has a distribution with the property that $E g'(\varepsilon) = 0$.

In our unit root problem \hat{c} solves $\sum \psi(\Delta y_t - \hat{c} y_{t-1}/T) y_{t-1} = 0$, which suggests

the identification condition

$$0 = \mathbb{E} \sum y_{t-1} \mathbb{E}_{t-1} \psi(\varepsilon_t) \Rightarrow \text{Identification condition is } \mathbb{E} \psi(\varepsilon_t) = 0.$$

For the LAD problem $\psi(\varepsilon_t) = \text{sign}(\varepsilon_t)$ and the identification condition is $\mathbb{E} \text{sign}(\varepsilon_t) = 0$. The condition holds for LAD only if the errors have zero medians.

What assumptions do we need to ensure that $\mathbb{E} \psi(\varepsilon_t) = 0$?

1. Expression (2) demonstrates that $\mathbb{E} \psi(\varepsilon_t) = 0$ under correct specification (so $f = g$).
2. $\mathbb{E} \psi(\varepsilon_t)$ equals zero for the Gaussian likelihood, no matter what the distribution of the errors. The Gaussian likelihood has $g(x) = x^2/2$ and $\psi(x) = x$, so the assumption $\mathbb{E} \varepsilon_1 = 0$ insures that $\mathbb{E} \psi(\varepsilon_1) = 0$.
3. $\mathbb{E} \psi(\varepsilon_t)$ equals zero when $f \neq g$ and both functions are symmetric around zero.
4. When $f \neq g$ and f is not symmetric, $\mathbb{E} \psi(\varepsilon_t)$ can be different from zero. For example, for LAD estimation applied to the errors $\varepsilon_t = Z_t^2 - 1$, $\mathbb{E} \psi(\varepsilon_t) = \mathbb{E} \text{sign}(Z_t^2 - 1) \approx -.3656$.

It turns out that the optimal tests are not robust to unknown, asymmetric error distributions. To get the tests to work, we need to assume either that we know the distribution of ε , or we need to assume that ε comes from a symmetric distribution. Thus the optimal tests are not robust to unknown asymmetric error distributions. The Gaussian tests of Elliott et al. (1996) are the one exception - those tests are valid under fairly general forms of misspecification, including asymmetric errors.

Figure 2 depicts an example where the optimal tests reject a true null hypothesis too often. This is generally a problem with asymmetric errors.

Proposition 1 *Suppose that g is three times differentiable with bounded third derivatives, and suppose that the errors satisfy assumption (1). If $\mathbb{E} \psi(\varepsilon_t) \neq 0$ then in large samples tests based on \hat{c} , \hat{t} , \hat{l} and the Neyman-Pearson statistic all reject a true null hypothesis with probability approaching .5, no matter what the nominal size of the test. The M-tests based on \tilde{c} and \tilde{t} have the same limiting representations as in theorem 1.⁵ In large samples the M-tests have accurate size.*

M-tests are robust to asymmetric error densities. Let η denote the parameter that solves the equation $\mathbb{E} \psi(\varepsilon_t - \eta) = 0$. So for LAD estimation, $\psi(\varepsilon_t - \eta) = \text{sign}(\varepsilon_t - \eta)$ and η is the median of the errors. The M-estimator objective function can be rewritten

$$\sum g\left(\Delta y_t - \frac{c}{T} y_{t-1} - a\right) = \sum g\left(\tilde{\varepsilon}_t - \frac{c}{T} u_{t-1} - \left(a - \eta - b_1 \frac{c}{T}\right)\right),$$

⁵The M-tests have the same limiting representations as in theorem 1, with the nuisance parameters σ_ψ^2 , ρ , and ω replaced by $\text{Var}[\psi(\varepsilon_t - \eta)]$, $\text{Corr}[\varepsilon_t, \psi(\varepsilon_t - \eta)]$, and $-\int_{\mathbb{R}} \psi(x - \eta) df(x)$, where η denotes the parameter that solves $\mathbb{E} \psi(\varepsilon_t - \eta) = 0$.

with $\tilde{\varepsilon}_t = \varepsilon_t - \eta$. These recentered errors satisfy $\mathbb{E} \psi(\tilde{\varepsilon}_t) = 0$. Thompson (2001b) has shown that if $\mathbb{E} \psi(\varepsilon_t) \neq 0$ then $\tilde{a} \rightarrow \eta$ in probability and \tilde{c} has the same limiting distribution as in theorem 1, with a slight redefinition of the nuisance parameters (see footnote 5). Thus estimation of the free parameter a causes a power loss under correct specification but ensures robustness against incorrect specification.

We can avoid these centering problems by assuming that $\mathbb{E} \psi(\varepsilon_t) = 0$. For example for LAD estimation we could assume that the median of ε is zero and leave the mean unspecified. The zero mean assumption is essential for nearly integrated models because it identifies the trend. If the mean is not zero, then the trends behave very differently under the unit root null than for stationary alternatives. Since u_t follows the process $\Delta u_t = \gamma u_{t-1} + \varepsilon_t$, we have

$$y_t = \beta_1 + \mathbb{E}(\varepsilon_1) \sum_{i=0}^{t-1} (\gamma + 1)^i + \sum_{i=0}^{t-1} (\gamma + 1)^i \tilde{\varepsilon}_{t-i}$$

with $\tilde{\varepsilon}_i = \varepsilon_i - \mathbb{E}(\varepsilon_1)$. If $\gamma = 0$, then y_t has both a unit root and a nonstochastic trend. If $\gamma < 0$, then y_t is stationary with the long run mean $\beta + \mathbb{E}(\varepsilon_1) \sum_{i=0}^{\infty} (\gamma + 1)^i$. Thus the zero mean assumption is essential if we wish to test for mean reversion around an intercept. Once we assume zero means, adding additional centering assumptions like zero medians takes us closer to assuming symmetric errors.

Sketch of proof of proposition 1.

When $\mathbb{E} \psi(\varepsilon_t) \neq 0$, the Neyman-Pearson statistic is not stochastically bounded. To understand why, notice that the statistic $T^{-1} \sum \psi(\varepsilon_t) u_{t-1}$ appearing in the approximation to the Neyman-Pearson statistic is not stochastically bounded. Lemma 3.1 of Phillips (1988) implies that we must divide by $T^{1/2}$ to get a limiting distribution:

$$\frac{1}{T^{3/2}} \sum \psi(\varepsilon_t) u_{t-1} \Rightarrow N(0, \sigma^2(c)).$$

where $\sigma^2(c) = (\sigma_\varepsilon \mathbb{E} \psi(\varepsilon_t) / c)^2 [1 + (e^{2c} - 1)/(2c) - 2(e^c - 1)/c]$. If g is three times differentiable with bounded third derivatives, the Neyman-Pearson statistic must also be divided by $T^{1/2}$:

$$\begin{aligned} T^{-1/2} \{L(\bar{c}, 0) - L(0, 0)\} &= -\frac{\bar{c}}{T^{3/2}} \sum \psi(\varepsilon_t) u_{t-1} + (\bar{c}^2 - 2\bar{c}c) \frac{\omega}{2T^{5/2}} \sum u_{t-1}^2 + o_p(1) \\ &= -\frac{\bar{c}}{T^{3/2}} \sum \psi(\varepsilon_t) u_{t-1} + o_p(1). \end{aligned}$$

Suppose we form the Neyman-Pearson statistic and use the critical value $q(\bar{c})$ constructed under the assumption that $\mathbb{E} \psi(\varepsilon_t) = 0$. If $\mathbb{E} \psi(\varepsilon_t) \neq 0$ then under the null hypothesis the probability of rejecting is

$$\lim_{T \rightarrow \infty} \Pr[L(\bar{c}, 0) - L(0, 0) < q(\bar{c})] = \Pr[N(0, \bar{c}^2 \sigma^2(0)) < 0] = .5,$$

where $\sigma^2(0) = \lim_{c \rightarrow 0} \sigma^2(c) = (\sigma_\varepsilon \mathbb{E} \psi(\varepsilon_t))^2 / 3$. In large samples the Neyman-Pearson test rejects a true null hypothesis 50% of the time, no matter what the nominal size of the test.

A proof by contradiction shows that error misspecification may also cause the constrained maximum likelihood estimator to be stochastically unbounded. Suppose that $c = \beta_1 = 0$ and g is three times differentiable with bounded third derivatives. If \hat{c} is stochastically bounded, the minimized objective function admits the approximation

$$L(\hat{c}, 0) - \sum g(\varepsilon_t) = -\hat{c}T^{-1} \sum \psi(\varepsilon_t) u_{t-1} + \frac{\hat{c}}{2}T^{-2} \sum g''(\varepsilon_t) u_{t-1}^2 + o_p(1).$$

If \hat{c} is stochastically bounded, then in large samples \hat{c} must converge to the minimizer of $L(\hat{c}, 0) - \sum g(\varepsilon_t)$. The minimizer $[T^{-2} \sum g''(\varepsilon_t) u_{t-1}^2]^{-1} T^{-1} \sum \psi(\varepsilon_t) u_{t-1}$ is not stochastically bounded and we have our contradiction. ■

2.3 Some Monte Carlo results

A Monte Carlo study demonstrates the size distortions that occur under incorrect specification. Table 1 presents rejection frequencies for ten tests under various assumptions about the true data generating process. The abbreviations in the table are:

1. ERS - the DF-GLS test of Elliott et al. (1996). This test efficiently handles the intercept for Gaussian errors but does not use the information in thick-tailed error distributions.
2. Adap - the adaptive test of Shin and So (1999). This test adapts to the error distribution but does not efficiently handle the intercept.
3. \tilde{c} test, LAD - Thompson's (2001a) version of the test based on the LAD M-estimator.⁶ The test is asymptotically equivalent to the test based on \tilde{c} , and in some cases has more accurate size. This test does not efficiently handle the intercept.
4. \tilde{c} test, t3 - Thompson's (2001a) version of the test based on the Student's t3 M-estimator.
5. Trend-optimal LAD NP, \hat{c} , and \hat{t} - these tests are optimal for a double exponential likelihood. NP denotes the Neyman-Pearson test statistic evaluated at $\bar{c} = -3$ and $\beta_1 = 0$: $L(-3, 0) - L(0, 0)$. The three tests efficiently handle the trend and will be more powerful than the DF-GLS test for many thick-tailed error distributions. When the errors come from an asymmetric distribution the tests will over reject a true null hypothesis.
6. Trend-optimal t3 NP, \hat{c} , and \hat{t} - the optimal tests for a Student's t3 likelihood with $\beta_1 = 0$. These tests are not robust to asymmetric errors.

⁶The test rejects for small values of $\tilde{c} = \left[\sum (y_{t-1} - \bar{y})^2 \right]^{-1} [T \sum (y_{t-1} - \bar{y}) \psi(\hat{\varepsilon}_t)]$, where $\hat{\varepsilon}_t = \Delta y_t - \tilde{a}_{1,R}$ and $\tilde{a}_{1,R} = \operatorname{argmin}_{a_1} \sum g(\Delta y_t - a_1)$. Thompson (2001a) shows that $|\omega\tilde{c} - \tilde{c}| \xrightarrow{p} 0$.

The bold numbers in table 1 illustrate the size problems with asymmetric errors.⁷ All of the tests have reasonably accurate sizes for the symmetric Student's t4 errors. However the asymmetric log normal and chi-squared errors cause the trend-optimal tests to over-reject true null hypotheses, and the problem gets worse as the sample size grows from $T = 100$ to $T = 1000$. Proposition 1 predicts that as the sample size grows the trend-optimal tests will reject a true null hypothesis with probability approaching .50. The Monte Carlo results seem to confirm the prediction, as the rejection frequencies for samples of 1000 are close to .50. The ERS, adaptive and \tilde{c} tests have accurate sizes for the asymmetric distributions.

The results demonstrate that the ERS test is a viable alternative to the robust tests even when the errors are not Gaussian. The ERS test has accurate size and good power for all four error distributions. Somewhat surprisingly, the ERS test even has accurate size for the infinite variance Cauchy distribution. The adaptive test works well for the asymmetric log normal and chi-squared distributions, but has poor power for the Cauchy errors. The \tilde{c} tests have accurate size for all four distributions. For the Cauchy errors the \tilde{c} tests are very powerful.

3 Optimal tests with a time trend

If the regressors include a linear time trend,

$$L(c, \beta_1, \beta_2) = \sum_{t=2}^T g \left(\Delta y_t - \frac{c}{T} y_{t-1} + \beta_1 \frac{c}{T} - \beta_2 \left(1 - (t-1) \frac{c}{T} \right) \right)$$

is the log-likelihood function conditional on the first observation. Suppose we form the Neyman-Pearson statistic $L(\bar{c}, \beta_1, \beta_2) - L(0, \beta_1, \beta_2)$ with the unknown coefficients replaced by the guess $b = (b_1, b_2)'$:

$$\begin{aligned} L(\bar{c}, b_1, b_2) - L(0, b_1, b_2) &= \sum g \left(\varepsilon_t - \frac{\bar{c} - c}{T} u_{t-1} + (b_1 - \beta_1) \frac{\bar{c}}{T} - (b_2 - \beta_2) \left(1 - (t-1) \frac{\bar{c}}{T} \right) \right) \\ &\quad - \sum g \left(\varepsilon_t + \frac{c}{T} u_{t-1} - (b_2 - \beta_2) \right). \end{aligned}$$

In large samples the term $(b_1 - \beta_1) \bar{c}/T$ disappears from this expression, so the guess b_1 does not matter. The terms $(b_2 - \beta_2) (1 - (t-1) \bar{c}/T)$ and $(b_2 - \beta_2)$ do not disappear,

⁷Critical values for first four tests are obtained using the methods described in Elliott et al. (1996), Shin and So (1999) and Thompson (2001a). In all cases the errors are iid and no correction is made for serial correlation. Critical values for the trend-optimal tests are obtained by simulating from the asymptotic distributions in theorem 1. The representations depend on the nuisance parameters σ_ε , σ_ψ , ρ , and ω . For all four tests the nuisance parameters are estimated using the formulas $\hat{\sigma}_\varepsilon^2 = T^{-1} \sum (\hat{\varepsilon}_t - \bar{\varepsilon})^2$, $\hat{\sigma}_\psi^2 = T^{-1} \sum (\psi(\hat{\varepsilon}_t) - \bar{\psi})^2$ and $\hat{\rho} = (\hat{\sigma}_\varepsilon \hat{\sigma}_\psi)^{-1} T^{-1} \sum (\hat{\varepsilon}_t - \bar{\varepsilon}) \psi(\hat{\varepsilon}_t)$, where $\hat{\varepsilon}_t$ is a residual and $\bar{\varepsilon}$ and $\bar{\psi}$ are sample averages. For the t3 estimator $\omega = E \psi'(\varepsilon_t)$ is estimated by $T^{-1} \sum \psi'(\hat{\varepsilon}_t)$. For the LAD test $\omega = 2f(\eta)$, which is estimated by the usual kernel estimator of the density of $\hat{\varepsilon}_t$ evaluated at zero: $\hat{f}(\eta) = (hT)^{-1} \sum \phi(\hat{\varepsilon}_t/h)$, where ϕ is the density function of a standard normal variable and h is the bandwidth $1.06 \hat{\sigma}_\varepsilon T^{-1/5}$. For the Neyman-Pearson test c is not estimated and there is no residual, so we use the nuisance parameters computed for the \tilde{c} estimator.

Table 1: Rejection frequencies for selected tests in the model with no time trend. The boldface numbers show the size distortions from using trend-optimal tests with asymmetric errors.

c	T	γ			\hat{c} tests		Trend-optimal LAD			Trend-optimal t3		
			ERS	Adap	LAD	t3	NP	\hat{c}	\hat{t}	NP	\hat{c}	\hat{t}
Student's t errors, 4 degrees of freedom												
0	100	0	.076	.041	.044	.044	.035	.034	.025	.054	.050	.054
-5	100	-.05	.426	.194	.234	.248	.257	.281	.230	.461	.414	.469
-10	100	-.1	.839	.477	.568	.625	.381	.682	.528	.706	.826	.816
0	1000	0	.051	.047	.051	.051	.042	.042	.035	.050	.049	.050
Cauchy errors												
0	100	0	.042	.010	.043	.043	.001	.001	.000	.004	.004	.003
-5	100	-.05	.285	.147	.968	.933	.973	.938	.814	.970	.996	.965
-10	100	-.1	.884	.206	.998	.923	.993	1.00	.991	.942	1.00	.997
0	1000	0	.030	.000	.044	.044	.000	.000	.000	.000	.000	.000
Log normal errors, centered to have zero mean and unit variance												
0	100	0	.065	.020	.039	.037	.395	.153	.401	.277	.113	.316
-5	100	-.05	.396	.577	.452	.470	.637	.691	.682	.775	.737	.763
-10	100	-.1	.826	.893	.835	.871	.801	.916	.887	.966	.973	.970
0	1000	0	.050	.031	.049	.050	.494	.485	.494	.485	.418	.483
Chi-squared errors, centered to have zero mean and unit variance												
0	100	0	.070	.018	.040	.037	.447	.191	.458	.331	.120	.352
-5	100	-.05	.409	.699	.354	.383	.623	.677	.682	.766	.707	.749
-10	100	-.1	.831	.929	.769	.841	.756	.897	.870	.958	.966	.962
0	1000	0	.053	.029	.053	.053	.498	.486	.497	.493	.405	.488

Notes: The initial condition is $u_0 = 0$. The trend coefficients are $(\beta_0, \beta_1) = (1, 0)$. Critical values are calculated by the method described in footnote 7. There are 20000 Monte Carlo repetitions.

so unless we know the true β_2 we cannot obtain the power bound $\Pi(c)$ derived in the last section. It is important to come up with a good guess for β_2 .

In most situations the trend parameter β_2 is unrelated to the unit root testing problem. Following Dufour and King (1991) and Elliott et al. (1996), it is natural to restrict attention to the family of tests which are invariant to the value of β_2 . By the well-known result of Lehmann (1959) (page 249), the most powerful invariant test of the hypothesis $c = 0$ versus the alternative $c = \bar{c}$ rejects for large values of

$$\int_{-\infty}^{\infty} \exp \{-L(\bar{c}, \beta_1, b)\} db / \int_{-\infty}^{\infty} \exp \{-L(0, \beta_1, b)\} db.$$

Elliott et al. (1996) encountered a similar integral for the Gaussian likelihood, where $g(x) = x^2/2$. In the Gaussian case $L(\bar{c}, \beta_1, b)$ is quadratic in b and the method of “completing the square” leads to a closed form solution for the integral. Since for many non Gaussian likelihoods it is not obvious how to solve this integral, I

approximate $L(\bar{c}, \beta_1, b)$ with a quadratic function of b and show that the approximate solution is asymptotically equivalent to the exact solution. This approach is a variant of Laplace's method (see Judd (1998) page 525). Laplace used this approach to approximate a similar integral over the double exponential distribution.⁸

The quadratic approximation is

$$\begin{aligned}\mathcal{Q}(\bar{c}, \phi) &= -\sum \psi(\varepsilon_t) z_t(\bar{c}, \phi) + \frac{\omega}{2} \sum z_t^2(\bar{c}, \phi) \\ z_t(\bar{c}, \phi) &= \frac{\bar{c} - c}{T} u_{t-1} + \frac{\phi}{\sqrt{T}} \left(1 - \bar{c} \frac{t-1}{T}\right),\end{aligned}$$

$\phi = \sqrt{T}(b - \beta_2)$. In lemma 1 in the appendix, I show that if $E\psi(\varepsilon_t) = 0$ then

$$L(\bar{c}, 0, \beta_2 + T^{-1/2}\phi) = \sum g(\varepsilon_t) + \mathcal{Q}(\bar{c}, \phi) + o_p(1).$$

In large samples the intercept β_1 disappears from the likelihood. In the proof of theorem 2 in the appendix I show that

$$\frac{\int \exp\{-L(\bar{c}, \beta_1, b_2)\} db_2}{\int \exp\{-L(0, \beta_1, b_2)\} db_2} = \frac{\int \exp\{-\mathcal{Q}(\bar{c}, \phi)\} d\phi + o_p(1)}{\int \exp\{-\mathcal{Q}(0, \phi)\} d\phi + o_p(1)}.$$

These integrals admit analytic solutions. Tedious algebraic manipulations lead to the result that the log of this ratio is asymptotically equivalent to

$$\min_{\phi} \mathcal{Q}(0, \phi) - \min_{\phi} \mathcal{Q}(\bar{c}, \phi)$$

plus terms that do not depend on c . This suggests the following theorem.

Theorem 2 *If $E\psi(\varepsilon_t) = 0$ and assumptions (1) and (2) hold, the most powerful invariant test is asymptotically equivalent to the test that rejects for small values of*

$$\min_b L(\bar{c}, 0, b) - \min_b L(0, 0, b).$$

Let $\pi^\tau(c, \bar{c})$ denote the limiting power function for the best invariant test indexed by \bar{c} when the true value of the locally autoregressive parameter is c :

$$\pi^\tau(c, \bar{c}) = \lim_{T \rightarrow \infty} \Pr \left[\min_{b_2} L(\bar{c}, 0, b) - \min_{b_2} L(0, 0, b) < q^\tau(\bar{c}) \right],$$

where $q^\tau(\bar{c})$ satisfies $\pi^\tau(0, \bar{c}) = \alpha$. The most powerful invariant test against the alternative c has power equal to $\Pi^\tau(c) \equiv \pi^\tau(c, c)$, the envelope power function.

Consider two estimators for c :

1. $(\tilde{c}, \tilde{a}_1, \tilde{a}_2) = \operatorname{argmin}_{(c, a_1, a_2)} \sum g(\Delta y_t - c y_{t-1}/T - a_1 - a_2(t-1)/T)$, with $a_1 = \beta_2 - \beta_1 c/T$ and $a_2 = -c\beta_2$. These are the usual M-estimators studied by Lucas (1995), Thompson (2001b), and Hasan and Koenker (1997) (see footnote 2).

⁸I thank Gary Chamberlain for making me aware of the links between Laplace's integration problem and this one.

2. $(\widehat{c}, \widehat{\beta}_2) = \operatorname{argmin}_{(c, \beta_2)} \sum g(\Delta y_t - cy_{t-1}/T - \beta_2 + c\beta_2(t-1)/T)$. This estimator was suggested by Xiao (2001).

In large samples $\beta_1 c/T$ is close to zero. This implies that the three parameters c , a_1 and a_2 can be written as just two, since $\lim_{T \rightarrow \infty} ca_1 = -a_2$. Under correct specification of the errors, \widehat{c} exploits the parameter restriction and a test that rejects for small \widehat{c} dominates one that rejects for small \widetilde{c} . In fact, the test based on \widehat{c} is asymptotically admissible, because its limiting power function touches the power envelope Π^τ . The test based on \widetilde{c} is not asymptotically admissible.

Another interesting test is the t-test based on the M-estimator, which rejects for small values of $\widehat{t} = [\sum \widehat{r}_{t-1}^2]^{1/2} \widetilde{c}$, where \widehat{r}_{t-1} is the residual from a least squares regression of y_{t-1} on $(1, t/T)$. This test is not asymptotically admissible, and is dominated by the constrained t and likelihood ratio tests, which reject for small values of

$$\widehat{t} = \sqrt{T^{-2} \sum \left(y_{t-1} - (t-1)\widehat{\beta}_2 \right)^2} \widehat{c} \quad \text{and} \quad \widehat{l} = -2 \left[\min_{c,b} L(c, 0, b) - \min_b L(0, 0, b) \right].$$

3.1 Asymptotic power functions

In order to derive the power functions of the various test statistics, it will prove useful to provide a limiting representation for the objective function. By Lemma 3.1 of Phillips (1988),

$$\mathcal{Q}(\bar{c}, \phi) \Rightarrow \mathcal{Q}^A(\bar{c}, \phi) \equiv -\sigma_\psi \int_0^1 P_{\bar{c}, \phi}(r) dS_\rho(r) + \frac{\omega}{2} \int_0^1 P_{\bar{c}, \phi}^2(r) dr,$$

where $P_{\bar{c}, \phi}(r) = \sigma_\varepsilon (\bar{c} - c) W_c(r) + \phi(1 - \bar{c}r)$ is a stochastic process. The following theorem is proven in the Appendix.

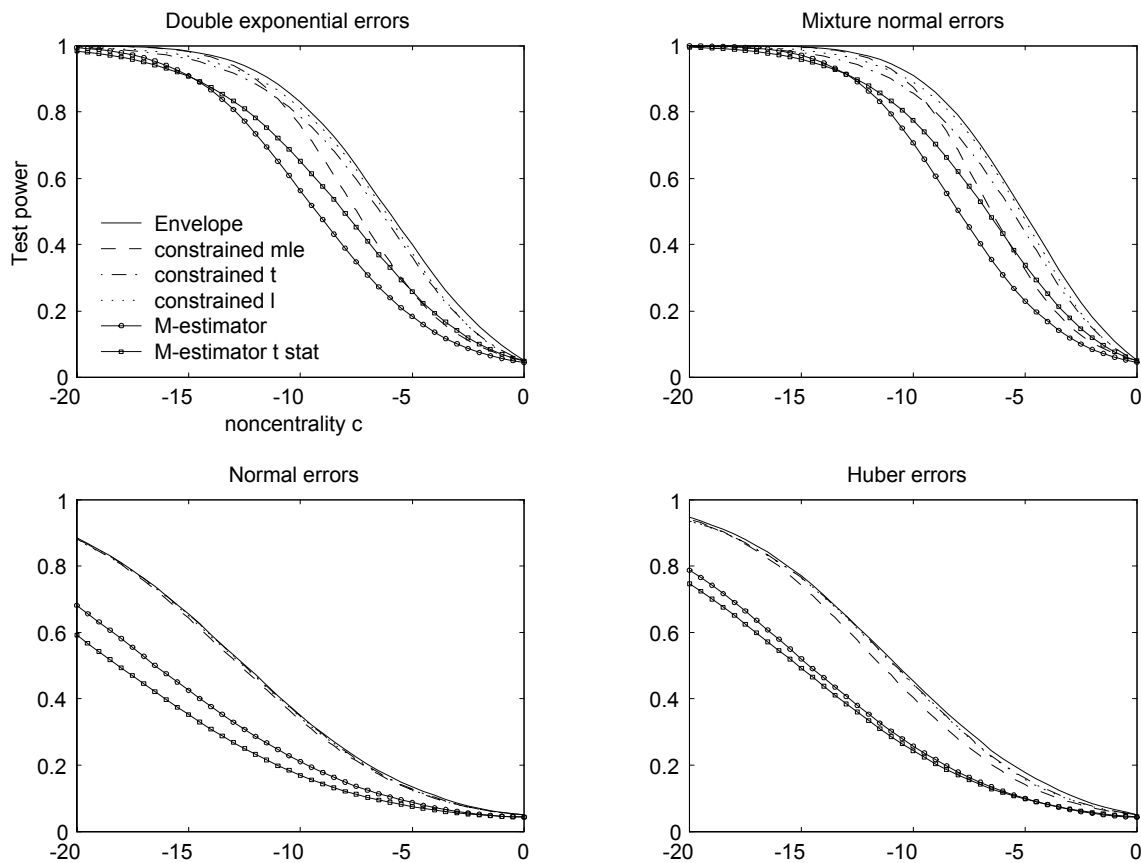
Theorem 3 *If $E\psi(\varepsilon_t) = 0$ and assumptions (1) and (2) hold, then*

1. $\min_b L(\bar{c}, 0, b) - \min_b L(0, 0, b) \Rightarrow \min_\phi \mathcal{Q}^A(\bar{c}, \phi) - \min_\phi \mathcal{Q}^A(0, \phi)$,
2. $(\widehat{c}, \widehat{\phi}) \Rightarrow (\widehat{\mathcal{C}}, \widehat{\mathcal{B}}) = \operatorname{argmin}_{(\mathcal{C}, \mathcal{B})} \mathcal{Q}^A(\mathcal{C}, \mathcal{B})$,
3. $\widehat{t} \Rightarrow \sigma_\varepsilon \widehat{\mathcal{C}} \sqrt{\int \left(W_c(t) - \widehat{\mathcal{B}}t \right)^2 dt}$,
4. $\widehat{l} \Rightarrow -2 \left[\min_{(\mathcal{C}, \mathcal{B})} \mathcal{Q}^A(\mathcal{C}, \mathcal{B}) - \min_{\mathcal{B}} \mathcal{Q}^A(0, \mathcal{B}) \right]$,
5. $\widetilde{c} \Rightarrow \sigma_\psi \left[\omega \sigma_\varepsilon \int D_c^2 \right]^{-1} \int D_c dS_\rho + c$, where $D_c(r) = W_c(r) - 2 \int_0^1 (2 - 3s - r(3 - 6s)) W_c(s) ds$,
6. $\widetilde{t} \Rightarrow (\sigma_\psi / \omega) \left[\int D_c^2 \right]^{-1/2} \int D_c dS_\rho + c \sigma_\varepsilon \left[\int D_c^2 \right]^{1/2}$.

Appendix A.2 provides a closed form expression for $\min_{\phi} \mathcal{Q}^A(\bar{c}, \phi) - \min_{\phi} \mathcal{Q}^A(0, \phi)$ in terms of stochastic integrals. Since $\mathcal{Q}^A(\bar{c}, \phi)$ is a nonlinear function of \bar{c} and ϕ , the asymptotic representation for \hat{c} does not admit an analytic solution in terms of random integrals. Appendix A.2 provides a method for simulating from the asymptotic distribution of \hat{c} , \hat{t} and \hat{l} .

Figure 3 plots the limiting power functions for the various tests in the model with a time trend. The curves are lower than the corresponding power envelopes for the model with an intercept only. Power rises as the tails of the error distribution become thicker; for Gaussian errors 50% power is achieved at -12.5 and for double exponential errors 50% power is achieved at about -6.0 .

Figure 3: Asymptotic power curves for unit root tests in the model with a time trend ($x_t = (1, t)$). The curves are drawn under the assumption of correct specification, so the g function used to form the test statistics is equal to the negative log-density of the errors f .



Notes: See notes to figure 1.

Figure 3 shows that the constrained \hat{c} and \hat{l} tests are asymptotically admissible. Careful examination of the figure leads to the conclusion that the constrained \hat{t} statistic is not asymptotically admissible. Neither M-test is asymptotically admissible. The \hat{c} test is point optimal when power is high, and the \hat{l} test is point optimal when power is close to one-half. Rothenberg (1984) describes similar results - he notes that in

standard (non unit root) models with no nuisance parameters, second order asymptotic theory predicts that estimator-based tests are optimal when power is high and likelihood ratio tests are optimal when power is close to 50%.

3.2 Failure of robustness to error misspecification

The analysis in the previous section was carried out under the assumption of correct specification, so $g = f$ where g is the function used to form the likelihood function and e^{-f} is the density of ε_t . As I discussed in section 2.2, correct specification insures that the centering condition $E\psi(\varepsilon_t) = 0$ holds. If $g \neq f$ then $E\psi(\varepsilon_t)$ may not equal zero. When $E\psi(\varepsilon_t) \neq 0$ the test statistics can behave badly.

Proposition 2 *Suppose that g is three times differentiable with bounded third derivatives, and suppose that the errors satisfy assumption (1). If $E\psi(\varepsilon_t) \neq 0$ then*

1. *Under the local alternative $\gamma = c/T$, $\hat{c} \xrightarrow{p} 0$ and $\hat{t} \xrightarrow{p} 0$. \hat{l} is $O_p(1)$ but does not have the distribution given in theorem 3. Therefore power against any local alternative approaches zero.*
2. *Under the fixed alternative $\gamma < 0$, $\hat{c} \xrightarrow{p} -\infty$, $\hat{t} \xrightarrow{p} -\infty$, and $\hat{l} \xrightarrow{p} +\infty$. Power against any fixed alternative approaches 1.*
3. *If $g''(x) > B > 0$ for all x , then under both local and fixed alternatives the best invariant test statistic $\min_{b \in K} L(\bar{c}, 0, b) - \min_{b \in K} L(0, 0, b) \xrightarrow{p} +\infty$, where K is a compact set. Power against any fixed or local alternative approaches zero.*
4. *The M-tests based on \tilde{c} and \tilde{t} have the same limiting distributions as in theorem 3.⁹*

Since the critical values for \hat{c} and \hat{t} are always negative, tests which reject for small \hat{c} and \hat{t} will have size converging to zero and power against any local alternative also converging to zero. The \hat{l} test has power equal to size for any local alternative, and its actual size will not match its nominal size, even in large samples. Since the critical values for the best invariant test are also negative, the best invariant test has size and power approaching zero against both fixed and local alternatives. The M-tests are robust to asymmetric errors. No matter what the error distribution, the M-tests have power against local alternatives and are consistent against fixed alternatives.

Thus none of the trend-optimal tests have power against local alternatives, but all except the best invariant test have power approaching 1 against fixed alternatives. In a large sample with a local alternative, the \hat{c} , \hat{t} and \hat{l} tests will be dominated by the robust M-tests. Furthermore, only the M-tests are useful for forming confidence intervals for the local parameter c , since that requires inverting a sequence of tests, each with power against local alternatives (see Elliott and Stock (2001) for the Gaussian case).

⁹The M-tests have the same limiting representations as in theorem 3, with the nuisance parameters redefined as in footnote 5.

On the other hand, the \widehat{c} and \widehat{t} tests have many desirable properties even when $\psi(\varepsilon_t) \neq 0$: in large samples they reject a true null hypothesis with probability less than any desired size, and they reject a fixed alternative $\gamma < 0$ with probability approaching 1. While the \widehat{t} test may get the size wrong, since the statistic is stochastically bounded both under the null and alternatives, the size distortions may be small. The magnitude of those distortions will be evaluated by Monte Carlo in section 3.3.

To understand the proposition, recall that the M-estimators $(\widetilde{c}, \widetilde{a}_1, \widetilde{a}_2)$ minimize the objective function

$$\sum g\left(\widetilde{\varepsilon}_t - \frac{\widetilde{c} - c}{T}u_{t-1} - \left(a_1 - \eta - \beta_2 + \beta_1 \frac{\widetilde{c}}{T}\right) - (a_2 + \widetilde{c}\beta_2) \frac{t-1}{T}\right)$$

where $\widetilde{\varepsilon}_t = \varepsilon_t - \eta$ and η denotes the parameter that solves $E\psi(\varepsilon_t - \eta) = 0$. If the condition $E\psi(\varepsilon_t) = 0$ fails to hold then $\widetilde{a}_1 \xrightarrow{p} \beta_2 + \eta$ and the “recentered” errors $\widetilde{\varepsilon}_t$ satisfy $E\psi(\widetilde{\varepsilon}_t) = 0$. Thus \widetilde{c} has the same limiting distribution as in theorem 3, with a slight redefinition of the nuisance parameters (see footnote 5). This result was shown by Thompson (2001*b*), and it implies that statement #4 of the proposition will hold. Since there is no free “recentering” parameter in the objective function for \widetilde{c} , the parameter on the time trend accomplishes the recentering. If $E\psi(\varepsilon_t) \neq 0$ then $\widehat{b}_2 \xrightarrow{p} \beta_2 + \eta$ and $\widehat{c} \xrightarrow{p} 0$, no matter what the local alternative c .

Sketch of proof of proposition 2.

To establish statement #1 of the proposition, consider the model with the local alternative $\gamma = c/T$. Define $\varphi = (\varphi_1, \varphi_2)' = (T^{1/2}\widetilde{c}, T^{1/2}(b_2 - \beta_2 - \eta))'$ and $m_t = T^{-1/2}(-\eta(t-1)/T, 1)'$. The likelihood function is

$$L(\varphi) = \sum g\left(\widetilde{\varepsilon}_t + \frac{c}{T}u_{t-1} - \beta_1 \frac{\varphi_1}{T^{3/2}} + \frac{\varphi_1\varphi_2}{T} \frac{t-1}{T} - m_t'\varphi\right).$$

A Taylor series expansion, combined with the usual asymptotic arguments (see Lemma 3.1 of Phillips (1988)), implies that

$$L(\varphi) - \sum g(\widetilde{\varepsilon}_t) = - \sum \widetilde{\psi}_t m_t'\varphi + \frac{\omega_\eta}{2} \varphi' \sum m_t m_t'\varphi + \frac{c\omega_\eta}{2T^2} \sum u_{t-1}^2 + R_T(\varphi),$$

where $\omega_\eta = E[\psi'(\widetilde{\varepsilon}_t)]$, $\widetilde{\psi}_t = \psi_t(\widetilde{\varepsilon}_t) + c\omega_\eta u_{t-1}/T$ and $\sup_{\varphi \in K} |R_T(\varphi)| \xrightarrow{p} 0$ for any compact set K . By the same argument used to prove lemma 2, φ is stochastically bounded. Therefore, by the argmax continuous mapping theorem of van der Vaart and Wellner (1996) (page 286), $\widehat{\varphi} = \operatorname{argmin}_\varphi L(\varphi)$ converges in probability to the minimizer of the approximating quadratic function, so

$$\widehat{\varphi} = \left[\omega_\eta \sum m_t m_t'\right]^{-1} \left[\sum m_t \widetilde{\psi}_t\right] + o_p(1).$$

$\sum m_t m_t'$ converges in probability to a nonrandom matrix, and $\sum m_t \widetilde{\psi}_t$ converges to a vector of mean-zero Gaussian random variables. In large samples $\widehat{\varphi}$ has a mean zero Gaussian distribution, which implies that $\widehat{c} \xrightarrow{p} 0$, $\widehat{t} \xrightarrow{p} 0$, and $b_2 \xrightarrow{p} \beta_2 + \eta$.

The distribution of \widehat{l} is obtained by substituting $\widehat{\varphi}$ back into the likelihood function. Under the local alternative $\gamma = c/T$, we obtain

$$\begin{aligned}\widehat{l} &= -2 \left[\min_{\varphi} L(\varphi) - \min_{\varphi, \varphi_1=0} L(\varphi) \right] + o_p(1) \\ &= \omega_{\eta}^{-1} \left\{ \left(T^{-1/2} \sum \widetilde{\psi}_t \right)^2 - \left(\sum \widetilde{\psi}_t m'_t \right) \left(\sum m_t m'_t \right)^{-1} \left(\sum m_t \widetilde{\psi}_t \right) \right\} + o_p(1).\end{aligned}$$

This is an $O_p(1)$ variable, but the limiting distribution differs from the one in theorem 1.

Statement #2 of the proposition says that the \widehat{c} , \widehat{t} and \widehat{l} tests are consistent against any fixed alternative $\gamma < 0$. The likelihood function evaluated at $(\overline{\gamma}, b)$ may be written

$$\sum g(\varepsilon_t - \eta - (\overline{\gamma} - \gamma) u_{t-1} + [(b_1 - \beta_1) \overline{\gamma} - \eta] - (b_2 - \beta_2) (1 - \overline{\gamma} (t-1))).$$

Since u_{t-1} is stationary under the fixed alternative, is it straightforward to show using Taylor series-based arguments that $\widehat{\gamma} = \widehat{c}/T$ is consistent for γ . Therefore $\widehat{c} \xrightarrow{p} -\infty$ and $\widehat{t} \xrightarrow{p} -\infty$, and an argument based on a Taylor series expansion demonstrates that $\widehat{l} \xrightarrow{p} +\infty$. I omit the proofs to save space.

To show statement #3 of the proposition, define the parameter η_{γ} which satisfies $E\psi(\varepsilon_t + \gamma u_{t-1} - \eta_{\gamma}) = 0$. For fixed $\gamma < 0$, $\varepsilon_t + \gamma u_{t-1}$ is a stationary random variable and the expectation exists. By a Taylor series expansion,

$$\frac{L(\overline{c}, 0, b_2)}{T} = \sum \frac{g(\varepsilon_t + \gamma u_{t-1} - \eta_{\gamma})}{T} + \sum \frac{\psi(\varepsilon_t + \gamma u_{t-1} - \eta_{\gamma})}{T} z_t^* + \sum \frac{\psi'(\varepsilon_t^*)}{2T} (z_t^*)^2,$$

where $z_t^* = -\overline{c}u_{t-1}/T - \beta_1\overline{c}/T + \eta_{\gamma} - (b_2 - \beta_2) \left(1 - \overline{c}\frac{t-1}{T}\right)$ and $|\varepsilon_t^* + \gamma u_{t-1} - \eta_{\gamma}| \leq |z_t^*|$. If $b_2 \in K$ then many of the terms are asymptotically negligible. We get the approximation

$$\frac{L(\overline{c}, 0, b_2)}{T} = \sum \frac{g(\varepsilon_t + \gamma u_{t-1} - \eta_{\gamma})}{T} + \sum \frac{\psi'(\varepsilon_t^*)}{2T} \left[\eta_{\gamma} - (b_2 - \beta_2) \left(1 - \overline{c}\frac{t-1}{T}\right) \right]^2 + o_p(1).$$

If $\overline{c} = 0$ this expression is minimized at $b_2 = \beta_2 + \eta_{\gamma}$, so $\min_{b \in K} T^{-1}L(0, 0, b) = T^{-1} \sum g(\varepsilon_t + \gamma u_{t-1} - \eta_{\gamma}) + o_p(1)$. If $\overline{c} \neq 0$, then since $g''(x) \geq B$ we have

$$\begin{aligned}\frac{L(\overline{c}, 0, b_2)}{T} &\geq \sum \frac{g(\varepsilon_t + \gamma u_{t-1} - \eta_{\gamma})}{T} + \frac{B}{2T} \sum \left[\eta_{\gamma} - (b_2 - \beta_2) \left(1 - \overline{c}\frac{t-1}{T}\right) \right]^2 + o_p(1) \\ &\geq \sum \frac{g(\varepsilon_t + \gamma u_{t-1} - \eta_{\gamma})}{T} + \frac{B\eta_{\gamma}^2\overline{c}^2}{24} + o_p(1).\end{aligned}$$

Therefore $T^{-1} \{ \min_{b \in K} L(\overline{c}, 0, b_2) - \min_{b \in K} L(\overline{c}, 0, b) \} \geq (24)^{-2} B\eta_{\gamma}^2\overline{c}^2 + o_p(1)$, and the best invariant test converges to $+\infty$ under any fixed alternative. Using the same arguments it is also possible to show that the best invariant test converges to $+\infty$ under any local value $\gamma = c/T$ (including the null $c = 0$). The proof is omitted to save space. ■

3.3 Some Monte Carlo results

Table 2 presents rejection frequencies for various tests in the model with a time trend. The tests are the trend versions of the tests that appeared in Table 1, except that the \hat{l} test appears in Table 2 in place of the \hat{t} test. I made this substitution because the \hat{t} test is not asymptotically admissible in the model with a time trend.¹⁰

The power losses from using the trend optimal \hat{c} and \hat{l} tests are small for samples of 100 observations, but get larger for samples of 1000. For the asymmetric log normal and chi squared error distributions, power against the local alternative $\bar{c} = -10$ declines as the sample size grows. Power against the fixed alternative $\gamma = -1$ increases with the sample size. This can be seen by comparing the results for the samples with $(T, \bar{c}) = (100, -10)$ against the samples with $(T, \bar{c}) = (1000, -100)$. In each case $\gamma = -1$. These results are consistent with proposition 2, which states that power against $\bar{c} = -10$ converges to zero as the sample grows and power against $\gamma = -1$ converges to one.

In the samples of 1000 observations, asymmetric errors cause the NP test to have low power against both the fixed and local alternative. This is consistent with proposition 2, which predicts that power against both kinds of alternatives converges to zero as the sample size grows.

No test dominates the others. The ERS test performs poorly for the Cauchy, log normal and chi squared errors. The adaptive test does well for the asymmetric distributions but has very low power with Cauchy errors. The \tilde{c} tests perform well for all the error distributions and sample sizes, but are generally dominated by the trend optimal \hat{c} and \hat{l} tests for samples of 100.

A Appendix

A.1 Proofs

In this appendix we prove theorems (1)-(3). Throughout the appendix it will prove useful to make use of the normalized likelihood $\mathcal{L}(\bar{c}, b_1, \phi) = L(\bar{c}, b_1, \beta_2 + T^{-1/2}\phi) - \sum g(\varepsilon_t)$. The remainder term $R_T(\bar{c}, b_1, \phi)$ is defined to be the difference between \mathcal{L} and its quadratic approximation \mathcal{Q} :

$$\mathcal{L}(\bar{c}, b_1, \phi) = \mathcal{Q}(\bar{c}, \phi) + R_T(\bar{c}, b_1, \phi).$$

A.1.1 Preliminary lemmas

Lemma 1 *Let K denote a compact set. If $\mathbb{E}\psi(\varepsilon_t) = 0$, and if assumptions (1) and (2) hold, then $\sup_{(\bar{c}, b_1, \phi) \in K} |R_T(\bar{c}, b_1, \phi)| \xrightarrow{p} 0$.*

¹⁰As was the case for Table 1, the \tilde{c} tests are Thompson's (2001a) versions of the tests. In the model with a time trend the test rejects for small values of $\tilde{c} = [\sum r_t^2]^{-1} [T \sum r_t \psi(\hat{\varepsilon}_t)]$, where r_t is the residual from a least squares regression of y_{t-1} on $(1, t)$, and $\hat{\varepsilon}_t = \Delta y_t - \tilde{a}_{1,R} - \tilde{a}_{2,R}t/T$ with $(\tilde{a}_{1,R}, \tilde{a}_{2,R}) = \operatorname{argmin}_{(a_1, a_2)} \sum g(\Delta y_t - a_1 - a_2t/T)$. Thompson (2001a) shows that $|\omega\tilde{c} - \tilde{c}| \xrightarrow{p} 0$.

Table 2: Rejection frequencies for selected tests in the model with a time trend. The boldface numbers illustrate that, with asymmetric errors, the trend optimal tests lose power against the local alternative $c = -10$ as the sample size increases.

c	T	γ	ERS	Adap	\tilde{c} tests		Trend-optimal LAD			Trend-optimal t3		
					LAD	t3	NP	\hat{c}	\hat{l}	NP	\hat{c}	\hat{l}
Student's t errors, 4 degrees of freedom												
0	100	0	.048	.023	.030	.028	.033	.030	.028	.068	.055	.061
-10	100	-1	.300	.180	.254	.279	.290	.306	.341	.555	.487	.561
-10	1000	-0.1	.294	.304	.341	.370	.338	.431	.429	.554	.544	.605
-100	1000	-1	1.00	.991	1.00	1.00	.910	1.00	1.00	1.00	1.00	1.00
Cauchy errors												
0	100	0	.024	.009	.038	.043	.001	.001	.001	.004	.006	.006
-10	100	-1	.181	.155	.984	.897	.988	.976	.991	.957	.999	.980
-10	1000	-0.1	.184	.000	1.00	.980	1.00	1.00	1.00	.991	1.00	.997
-100	1000	-1	.998	.000	1.00	.953	1.00	1.00	1.00	.766	1.00	.991
Log normal errors, centered to have zero mean and unit variance												
0	100	0	.036	.011	.028	.020	.022	.020	.019	.045	.045	.050
-10	100	-1	.275	.623	.577	.610	.457	.692	.619	.873	.811	.877
-10	1000	-0.1	.288	0.943	.803	.806	.004	.059	.078	.157	.502	.378
-100	1000	-1	1.00	.952	1.00	1.00	.027	1.00	1.00	.514	1.00	1.00
Chi-squared errors, centered to have zero mean and unit variance												
0	100	0	.043	.008	.030	.020	.032	.028	.030	.048	.055	.055
-10	100	-1	.290	.704	.477	.519	.288	.628	.480	.807	.765	.829
-10	1000	-0.1	.296	1.00	.568	.623	.000	.034	.034	.029	.359	.210
-100	1000	-1	1.00	1.00	1.00	1.00	.003	.999	.998	.197	1.00	1.00

Notes: The initial condition is $u_0 = 0$. The trend coefficients are $(\beta_0, \beta_1) = (1, 1)$. "NP" denotes the best invariant test evaluated at $\bar{c} = -6$, so the test statistic is $\min_b L(-6, 0, b) - \min_b L(0, 0, b)$. Critical values are calculated by the method described in footnote 7. There are 20000 Monte Carlo repetitions.

We will show that $R_T(\bar{c}, b_1, \phi) \xrightarrow{p} 0$ pointwise in (\bar{c}, b_1, ϕ) . If $\mathcal{L}(\bar{c}, b_1, \phi)$ were a convex function of (\bar{c}, b_1, ϕ) then pointwise convergence would imply uniform convergence over compact sets (this is shown in lemma 1 of Hjort and Pollard (1993)). However, even though $g(x)$ is convex in x , $\mathcal{L}(\bar{c}, b_1, \phi)$ is not a convex function of (\bar{c}, b_1, ϕ) because $g(\varepsilon_t + (b_1 - \beta_1)\bar{c}/T - z_t(\bar{c}, \phi))$ is a nonlinear function of the parameters.

A reparameterization allows us to restore the link between pointwise and uniform convergence. Let $w_t = T^{-1/2} (T^{-1/2}, T^{-1/2}u_{t-1}, 1, (t-1)/T)'$ and $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)' = (\bar{c}(\beta_1 - b_1), \bar{c} - c, \phi, -\phi\bar{c})'$. We have the reparameterized objective function and remainder term

$$\begin{aligned} \mathcal{L}(\bar{c}, b_1, \phi) &= \mathcal{L}_R(\theta) = \sum g(\varepsilon_t - w_t'\theta) - \sum g(\varepsilon_t). \\ \tilde{R}_T(\theta) &= \mathcal{L}_R(\theta) + \sum \psi(\varepsilon_t) w_t'\theta - \frac{\omega}{2} \sum (w_t'\theta)^2 \end{aligned}$$

Since g is convex and $\varepsilon_t - w_t'\theta$ is a linear function of θ , $\mathcal{L}_R(\theta)$ is a convex function of θ . Therefore if $\tilde{R}_T(\theta) \xrightarrow{p} 0$ pointwise in θ then the convergence is uniform for θ in a compact set. The relationship between the original remainder term and the reparameterized remainder is

$$R_T(\bar{c}, b_1, \phi) = \tilde{R}_T(\theta) + \tilde{R}_{T,2}(\theta)$$

$$\tilde{R}_{T,2}(\theta) = -\theta_1 T^{-1} \sum \psi(\varepsilon_t) + \frac{\omega}{2} \sum \left\{ \frac{\theta_1^2}{T^2} + 2\frac{\theta_1}{T} \left(\frac{\theta_2}{T} u_{t-1} + \frac{\theta_3}{T^{1/2}} + \theta_4 \frac{t-1}{T^{3/2}} \right) \right\}.$$

The equality holds so long as θ satisfies the constraint $\theta_4 = -\theta_3(\theta_2 + c)$. For any compact set $K \subset \mathbb{R}^3$, let $\tilde{K} \subset \mathbb{R}^4$ denote a compact set large enough so that if $\{\theta_1, \theta_2, \theta_3\} \in K$ then $\{\theta_1, \theta_2, \theta_3, -\theta_3(\theta_2 + c)\} \in \tilde{K}$. We now have a bound for the remainder R_T :

$$\sup_{(\bar{c}, b_1, \phi) \in K} |R_T(\bar{c}, b_1, \phi)| \leq \sup_{\theta \in \tilde{K}, \theta_4 = -\theta_3(\theta_2 + c)} \left| \tilde{R}_T(\theta) + \tilde{R}_{T,2}(\theta) \right| \leq \sup_{\theta \in \tilde{K}} \left| \tilde{R}_T(\theta) \right| + \sup_{\theta \in \tilde{K}} \left| \tilde{R}_{T,2}(\theta) \right|.$$

It is straightforward to show that $\sup_{\theta \in \tilde{K}} |\tilde{R}_{T,2}(\theta)| \rightarrow 0$ in probability for any compact set \tilde{K} . So to prove the theorem it is enough to show that $\tilde{R}_T(\theta) \xrightarrow{p} 0$ pointwise in θ . The pointwise convergence of $\tilde{R}_T(\theta)$ was proven in lemma 1 of Thompson (2001b). ■

Lemma 2 *In the model with a time trend, define $(\hat{c}, \hat{b}_2) = \operatorname{argmin}_{(\bar{c}, b_2)} L(\bar{c}, 0, b_2)$ and $\hat{\phi} = \sqrt{T}(\hat{b}_2 - \beta_2)$. If $E\psi(\varepsilon_t) = 0$, and if assumptions (1) and (2) hold, then \hat{c} and $\hat{\phi}$ are both stochastically bounded.*

If $\mathcal{L}(\bar{c}, 0, \phi)$ were convex in the parameters the argument in section 3 of Pollard (1991) could be used to show that the estimators $(\hat{c}, \hat{\phi})$ are stochastically bounded. However, even though $g(x)$ is convex in x , $\mathcal{L}(\bar{c}, 0, \phi)$ is not a convex function of (\bar{c}, ϕ) because $g(\varepsilon_t - \beta_1 \bar{c}/T - z_t(\bar{c}, \phi))$ is a nonlinear function of the parameters. We will pursue a related method of proof.

A reparameterization allows me to apply Pollard's (1991) arguments to this problem. Define $\theta = (\bar{c} - c, \phi + \beta_1 \bar{c}/T, -\phi \bar{c})'$, $w_t = T^{-1/2}(T^{-1/2}u_{t-1}, 1, (t-1)/T)'$, and

$$\mathcal{L}_R(\theta) = \mathcal{L}(\bar{c}, 0, \phi) = \sum g(\varepsilon_t - w_t'\theta).$$

Note that this reparameterization differs from the reparameterization in lemma 1 because here we take $b_1 = 0$. For some $k > 0$, define the compact set $K = \{(\bar{c}, \phi) : |\bar{c} - c| < k, |\phi + T^{-1/2}\beta_1 \bar{c}| < k\}$. For any $(\bar{c}, \phi) \notin K$, the corresponding θ vector is equal to vr where v is a vector with unit length and r is a scalar with $r > k$. Since g is convex and $\varepsilon_t - w_t'\theta$ is a linear function of θ , $\mathcal{L}_R(\theta)$ is a convex function of θ . So $k = (1 - \lambda)0 + \lambda r$ for $\lambda = k/r$, and by the convexity of $\mathcal{L}_R(\theta)$,

$$\mathcal{L}_R(vk) \leq (1 - \lambda)\mathcal{L}_R(0) + \lambda\mathcal{L}_R(vr), \text{ and } \mathcal{L}_R(0) = 0, \text{ which implies that } \mathcal{L}_R(vr) \geq \frac{r}{k}\mathcal{L}_R(vk).$$

In the proof of lemma 1 it was found that the approximation

$$\mathcal{L}_R(\theta) = - \sum \psi(\varepsilon_t) w_t' \theta + \frac{\omega}{2} \sum (w_t' \theta)^2 + o_p(1)$$

will hold uniformly over θ in a compact set. So for fixed k ,

$$\mathcal{L}_R(vr) \geq \frac{r}{k} \left[\frac{\omega k^2}{2} v' \left(\sum w_t w_t' \right) v - k \sum \psi(\varepsilon_t) w_t' v + o_p(1) \right].$$

By lemma 3.1 of Phillips (1988), $\sum w_t w_t'$ converges in distribution to a positive definite matrix with diagonal elements bounded away from zero with probability one. Therefore there exists $\epsilon > 0$ so that $\inf_{\|v\|=1} v' \left(\sum w_t w_t' \right) v > \epsilon$. So

$$\inf_{(\bar{c}, \phi) \notin K} \mathcal{L}(\bar{c}, 0, \phi) = \mathcal{L}_R(vr) \geq \left[\frac{k^2 \omega \epsilon}{2} - k \left\| \sum w_t \psi(\varepsilon_t) \right\| + o_p(1) \right].$$

Since $\left\| \sum w_t \psi(\varepsilon_t) \right\|$ is stochastically bounded (see Phillips's (1988) lemma 3.1), we can choose k large enough so that $k^2 \omega \epsilon / 2 - k \left\| \sum w_t \psi(\varepsilon_t) \right\| > \epsilon$ with probability arbitrarily close to 1. We have that in large samples,

$$\inf_{(\bar{c}, \phi) \notin K} \mathcal{L}(\bar{c}, 0, \phi) \geq \epsilon + o_p(1) > 0 = \mathcal{L}(0, 0, 0) \geq \inf_{\bar{c}, \phi} \mathcal{L}(\bar{c}, 0, \phi) = \mathcal{L}(\hat{c}, 0, \hat{\phi}).$$

So in large samples the estimators \hat{c} and $\hat{\phi}$ must be contained in K . Thus $(\hat{c}, \hat{\phi})$ are stochastically bounded and the theorem is proved. ■

A.1.2 Proofs of theorems

Proof of theorem 1.

Notice that $L(\bar{c}, 0) - L(0, 0) = \mathcal{L}(\bar{c}, 0, 0) - \mathcal{L}(0, 0, 0)$, and that $\hat{c} = \operatorname{argmin}_{\bar{c}} \mathcal{L}(\bar{c}, 0, 0)$. By lemma 1,

$$\mathcal{L}(\bar{c}, 0, 0) = -(\bar{c} - c) T^{-1} \sum \psi(\varepsilon_t) u_{t-1} + (\bar{c} - c)^2 \frac{\omega}{2} T^{-2} \sum u_{t-1}^2 + o_p(1).$$

The asymptotic representation for $L(\bar{c}, 0) - L(0, 0)$ follow from the following weak convergence result, which was proven by Phillips (1988) (see Lemma 3.1):

$$\left(T^{-1} \sum \psi(\varepsilon_t) u_{t-1}, T^{-2} \sum u_{t-1}^2 \right)' \Rightarrow \left(\sigma_\varepsilon \sigma_\psi \int W_c dS_\rho, \sigma_\varepsilon^2 \int W_c^2 \right)'$$

Since g is convex, $\mathcal{L}(\bar{c}, 0, 0)$ is convex in \bar{c} . By slightly modifying the argument in section 3 of Pollard (1991), it can be shown that the convexity of $\mathcal{L}(\bar{c}, 0, 0)$ implies that \hat{c} converges weakly to the minimizer of the quadratic approximation $\mathcal{Q}(\bar{c}, 0, 0)$, so

$$\hat{c} = \frac{T^{-1} \sum \psi(\varepsilon_t) u_{t-1}}{\omega T^{-2} \sum u_{t-1}^2} + c + o_p(1).$$

Therefore $\widehat{c} \Rightarrow \sigma_\psi [\omega \sigma_\varepsilon \int W_c^2]^{-1} \int W_c dS_\rho + c$, and the distribution of \widehat{t} and \widehat{l} follows similarly. The representations for \widetilde{c} and \widetilde{t} are provided in Theorem 1 of Thompson (2001b). ■

Proof of theorem 2.

To prove the theorem it is sufficient to show that

$$\int_{-\infty}^{\infty} \exp \{-\mathcal{L}(\bar{c}, \beta_1, \phi)\} d\phi = \int_{-\infty}^{\infty} \exp \{-\mathcal{Q}(\bar{c}, \phi)\} d\phi + o_p(1). \quad (3)$$

To see that this is indeed sufficient, notice that the best invariant test rejects for large values of

$$\frac{\int \exp \{-L(\bar{c}, \beta_1, b)\} db}{\int \exp \{-L(0, \beta_1, b)\} db} = \frac{\int \exp \{-L(\bar{c}, \beta_1, b) + \sum g(\varepsilon_t)\} db}{\int \exp \{-L(0, \beta_1, b) + \sum g(\varepsilon_t)\} db} = \frac{\int \exp \{-\mathcal{L}(\bar{c}, \beta_1, \phi)\} d\phi}{\int \exp \{-\mathcal{L}(0, \beta_1, \phi)\} d\phi},$$

where the last equality follows from the change of variables $\phi = \sqrt{T}(b - \beta_2)$. The discussion in section 3 indicates that if the approximation in 3 holds, then the best invariant test is asymptotically equivalent to rejecting for small values of

$$\min_{\phi} \mathcal{Q}(\bar{c}, \phi) - \min_{\phi} \mathcal{Q}(0, \phi). \quad (4)$$

This is asymptotically equivalent to $\min_b L(\bar{c}, 0, b) - \min_b L(0, 0, b)$. To see this, notice that

$$\min_b L(\bar{c}, 0, b) - \min_b L(0, 0, b) = \min_{\phi} \mathcal{L}(\bar{c}, 0, \phi) - \min_{\phi} \mathcal{L}(0, 0, \phi).$$

The convexity of g implies that for fixed \bar{c} , $\mathcal{L}(\bar{c}, 0, \phi)$ is convex in ϕ . By slightly modifying the method in section 3 of Pollard (1991), one can use the convexity to show that $\min_{\phi} \mathcal{L}(\bar{c}, 0, \phi)$ is asymptotically equal to $\min_{\phi} \mathcal{Q}(\bar{c}, \phi)$, the quadratic approximation given in lemma 1. So in large samples $\min_{\phi} \mathcal{L}(\bar{c}, 0, \phi) - \min_{\phi} \mathcal{L}(0, 0, \phi)$ is equivalent to the statistic in 4. Thus verifying the condition in 3 is sufficient to prove the theorem.

To verify 3 it will prove convenient to break the integral into two parts. For any positive k ,

$$\int_{-\infty}^{\infty} \exp \{-\mathcal{L}(\bar{c}, \beta_1, \phi)\} d\phi = \int_{-k}^k \exp \{-\mathcal{L}(\bar{c}, \beta_1, \phi)\} d\phi + \mathcal{I}^c(k, \bar{c}),$$

with $\mathcal{I}^c(k, \bar{c}) = \int_{\phi \notin [-k, k]} \exp \{-\mathcal{L}(\bar{c}, \beta_1, \phi)\} d\phi$. For any fixed k , lemma 1 implies that

$$\int_{-k}^k \exp \{-\mathcal{L}(\bar{c}, \beta_1, \phi)\} d\phi = \int_{-k}^k \exp \{-\mathcal{Q}(\bar{c}, \phi)\} d\phi + o_p(1).$$

The integral on the right hand side admits an analytic solution. Using that analytic solution is it straightforward to show that for all $\epsilon > 0$ we can pick k large enough so that

$$\lim_{T \rightarrow \infty} \Pr \left[\left| \int_{-k}^k \exp \{-\mathcal{L}(\bar{c}, \beta_1, \phi)\} d\phi - \int_{-\infty}^{\infty} \exp \{-\mathcal{Q}(\bar{c}, \phi)\} d\phi \right| < \epsilon \right] = 1.$$

It remains to show that $\mathcal{I}^c(k, \bar{c})$ is asymptotically negligible. Since $\mathcal{L}(\bar{c}, \beta_1, \phi)$ is convex in ϕ (for fixed \bar{c}), then if $\phi > k$ then $k = (1 - \lambda)0 + \lambda\phi$ and $\mathcal{L}(\bar{c}, \beta_1, k) \leq (1 - \lambda)\mathcal{L}(\bar{c}, \beta_1, 0) + \lambda\mathcal{L}(\bar{c}, \beta_1, \phi)$ with $\lambda = k/\phi$. Therefore

$$\text{(if } \phi > k \text{) then } \mathcal{L}(\bar{c}, \beta_1, \phi) \geq \frac{|\phi|}{k} [\mathcal{L}(\bar{c}, \beta_1, k) - \mathcal{L}(\bar{c}, \beta_1, 0)] + \mathcal{L}(\bar{c}, \beta_1, 0).$$

Similarly,

$$\text{if } \phi < -k \text{ then } \mathcal{L}(\bar{c}, \beta_1, \phi) \geq \frac{|\phi|}{k} [\mathcal{L}(\bar{c}, \beta_1, -k) - \mathcal{L}(\bar{c}, \beta_1, 0)] + \mathcal{L}(\bar{c}, \beta_1, 0).$$

By lemma 1, for fixed k we have

$$\begin{aligned} \mathcal{L}(\bar{c}, \beta_1, k) - \mathcal{L}(\bar{c}, \beta_1, 0) &= k^2 \left[\frac{\omega}{2T} \sum \left(1 - \bar{c} \frac{t-1}{T} \right)^2 \right] - k \left[\sum \frac{\psi(\varepsilon_t)}{T^{1/2}} \left(1 - \bar{c} \frac{t-1}{T} \right) \right] \\ &\quad + k \left[\frac{\omega(\bar{c} - c)}{2T^{3/2}} \sum u_{t-1} \left(1 - \bar{c} \frac{t-1}{T} \right) \right] + o_p(1). \end{aligned}$$

By the usual asymptotic arguments, $\lim_{T \rightarrow \infty} T^{-1} \sum (1 - \bar{c}(t-1)/T)^2 = \bar{c}^2/3 - \bar{c} + 1 \geq 1/4$ and the other terms are $O_p(1)$. So if $|\phi| > k$ then

$$\mathcal{L}(\bar{c}, \beta_1, \phi) \geq \frac{|\phi|}{k} [k^2/4 + kO_p(1) + o_p(1)] + O_p(1).$$

Plugging this bound into the integral, we obtain

$$\begin{aligned} \mathcal{I}^c(k, \bar{c}) &\leq \exp [O_p(1)] \int_{\phi \notin [-k, k]} \exp \left\{ -\frac{|\phi|}{k} [k^2/4 + kO_p(1) + o_p(1)] \right\} d\phi \\ &= \exp [O_p(1)] \frac{2k \exp \{ -[k^2/4 + kO_p(1) + o_p(1)] \}}{k^2/4 + kO_p(1) + o_p(1)}. \end{aligned}$$

Thus, for any $\epsilon > 0$, we can choose k large enough so that $\lim_{T \rightarrow \infty} \Pr [\mathcal{I}^c(k, \bar{c}) < \epsilon] = 1$. Thus the condition in 3 holds and the theorem is proven. ■

Proof of theorem 3.

In the proof of theorem 2, we showed that

$$\min_b L(\bar{c}, 0, b) - \min_b L(0, 0, b) = \min_\phi \mathcal{Q}(\bar{c}, \phi) - \min_\phi \mathcal{Q}(0, \phi) + o_p(1).$$

Lemma 3.1 of Phillips (1988) implies that $\mathcal{Q}(\bar{c}, \phi) \Rightarrow \mathcal{Q}^A(\bar{c}, \phi)$. Phillips's (1988) lemma also implies that $\text{argmin}_\phi \mathcal{Q}(\bar{c}, \phi)$ is stochastically bounded, and $\min_\phi \mathcal{Q}(\bar{c}, \phi) \Rightarrow \min_\phi \mathcal{Q}^A(\bar{c}, \phi)$ by the argmax continuous mapping theorem of van der Vaart and Wellner (1996) (page 286). We have derived the limiting representation for the best invariant test.

The argmax continuous mapping theorem also provides the limiting result for \hat{c} . Notice that $(\hat{c}, \hat{\phi}) = \text{argmin}_{\bar{c}, \phi} \mathcal{L}(\bar{c}, 0, \phi)$. Since, by lemma 2, \hat{c} and $\hat{\phi}$ are stochastically bounded, \hat{c} and $\hat{\phi}$ converge weakly to $(\hat{\mathcal{C}}, \hat{\mathcal{B}}) = \text{argmin}_{(\mathcal{C}, \mathcal{B})} \mathcal{Q}^A(\mathcal{C}, \mathcal{B})$. The limiting distributions of \hat{t} and \hat{l} follow from a similar argument.

Limiting representations for \tilde{c} and \tilde{t} are provided in Theorem 1 of Thompson (2001b). ■

A.2 Simulating the asymptotic distributions

Theorem 3 provides asymptotic representations for various test statistics in the model with a time trend. In this appendix we describe how to simulate from those distributions.

The best invariant test converges to $\min_{\phi} \mathcal{Q}^A(\bar{c}, \phi) - \min_{\phi} \mathcal{Q}^A(0, \phi)$. This is equal to

$$-\sigma_{\psi} \int \{V_c(t, \bar{c}) - V_c(t, 0)\} dS_{\rho}(t) + \frac{\omega}{2} \int \{V_c^2(t, \bar{c}) - V_c^2(t, 0)\} dt,$$

where

$$V_c(t, \bar{c}) = \sigma_{\varepsilon}(\bar{c} - c) W_c(s) + \frac{\int (1 - \bar{c}r) \{\sigma_{\psi} dS_{\rho}(r) - \sigma_{\varepsilon} \omega (\bar{c} - c) W_c(r) dr\} (1 - \bar{c}s)}{\omega (1 - \bar{c} + \bar{c}^2/3)}.$$

Simulating from this distribution is straightforward.

The normalized maximum likelihood estimators \hat{c} and $\hat{\phi} = \sqrt{T}(\hat{b}_2 - \beta_2)$ converge weakly to the random variables $\hat{\mathcal{C}}$ and $\hat{\mathcal{B}}$ which minimize the stochastic objective function $\mathcal{Q}^A(\mathcal{C}, \mathcal{B})$. I was unable to derive a simple expression for $\hat{\mathcal{C}}$ and $\hat{\mathcal{B}}$. Instead the variables are expressed implicitly as solutions to the minimization problem. Rewrite the objective function:

$$\begin{aligned} \mathcal{Q}^A(\mathcal{C}, \mathcal{B}) &= \sigma_{\varepsilon} A_0 + \mathcal{C} \sigma_{\varepsilon} A_1 + \mathcal{B} A_3 + \mathcal{B} \mathcal{C} A_4 \\ &+ \left[\frac{\lambda}{2} \sigma_{\varepsilon} \int W_c^2 \right] \mathcal{C}^2 + \frac{\lambda}{2\sigma_{\varepsilon}} \mathcal{B}^2 - \frac{\lambda}{2\sigma_{\varepsilon}} \mathcal{B}^2 \mathcal{C} - \left[\lambda \int r W_c \right] \mathcal{B} \mathcal{C}^2 + \frac{\lambda}{6\sigma_{\varepsilon}} \mathcal{B}^2 \mathcal{C}^2 \end{aligned}$$

where $\lambda = \omega \sigma_{\varepsilon} / \sigma_{\psi}$ and

$$\begin{aligned} A_0 &= \int W_c(r) dS_{\rho}(r) + \frac{\lambda}{2} \mathcal{C}^2 \int W_c^2 \\ A_1 &= - \int W_c(r) dS_{\rho}(r) - \lambda c \int W_c^2 \\ A_3 &= S_{\rho}(1) - \lambda c \int W_c(r) \\ A_4 &= - \int r dS_{\rho}(r) + \lambda \int W_c(r) + \lambda c \int r W_c(r). \end{aligned}$$

$\mathcal{Q}^A(\mathcal{C}, \mathcal{B})$ has at least one minimum. Take the derivatives of the function with respect to \mathcal{C} and \mathcal{B} :

$$\begin{aligned} \frac{\partial \mathcal{Q}^A(\mathcal{C}, \mathcal{B})}{\partial \mathcal{C}} &= \sigma_{\varepsilon} A_1 + \mathcal{B} A_4 + \left[\lambda \sigma_{\varepsilon} \int W_c^2 \right] \mathcal{C} - \frac{\lambda}{2\sigma_{\varepsilon}} \mathcal{B}^2 - \left[2\lambda \int r W_c \right] \mathcal{B} \mathcal{C} + \frac{\lambda}{3\sigma_{\varepsilon}} \mathcal{B}^2 \mathcal{C} \\ \frac{\partial \mathcal{Q}^A(\mathcal{C}, \mathcal{B})}{\partial \mathcal{B}} &= A_3 + \mathcal{C} A_4 + \frac{\lambda}{\sigma_{\varepsilon}} \mathcal{B} - \frac{\lambda}{\sigma_{\varepsilon}} \mathcal{B} \mathcal{C} - \left[\lambda \int r W_c \right] \mathcal{C}^2 + \frac{\lambda}{3\sigma_{\varepsilon}} \mathcal{B} \mathcal{C}^2 \end{aligned}$$

The values of \mathcal{C} and \mathcal{B} which minimize $\mathcal{Q}^A(\mathcal{C}, \mathcal{B})$ set the partial derivatives to zero. Solve $\partial \mathcal{Q}^A(\mathcal{C}, \mathcal{B}) / \partial \mathcal{B} = 0$ for \mathcal{B} to obtain

$$\mathcal{B}(\mathcal{C}) = \frac{\sigma_{\varepsilon} \left[\lambda \int r W_c \right] \mathcal{C}^2 - A_3 - \mathcal{C} A_4}{\lambda (1 - \mathcal{C} + \mathcal{C}^2/3)}.$$

Substitute the solution for $\mathcal{B}(\mathcal{C})$ into the equation $\partial \mathcal{Q}^A(\mathcal{C}, \mathcal{B}) / \partial \mathcal{C} = 0$ to show that $\hat{\mathcal{C}}$ is the root of a fifth order polynomial:

$$\begin{aligned}
0 = & [18\lambda A_1 - 18A_4 A_3 - 9A_3^3] \\
& + \mathcal{C} \left[6A_3^2 - 18A_4^2 - 36A_1\lambda + 18\lambda^2 \int W_c^2 + 36 \left(\int rW_c \right) \lambda A_3 \right] \\
& + \mathcal{C}^2 \left[9A_4^2 + 30A_1\lambda + 6A_3 A_4 - 36\lambda^2 \int W_c^2 + 18\lambda(3A_4 - A_3) \int rW_c \right] \\
& + \mathcal{C}^3 \left[30\lambda^2 \int W_c^2 - 36 \left(\int rW_c \right)^2 \lambda^2 - 12A_1\lambda - 36A_4\lambda \int rW_c \right] \\
& + \mathcal{C}^4 \left[27\lambda^2 \left(\int rW_c \right)^2 - 12\lambda^2 \int W_c^2 + 2A_1\lambda + 6A_4\lambda \int rW_c \right] \\
& + \mathcal{C}^5 \left[2\lambda^2 \int W_c^2 - 6 \left(\int rW_c \right)^2 \lambda^2 \right]
\end{aligned}$$

Notice that since A_0 , A_1 , A_3 and A_4 depend on ρ and λ and on no other nuisance parameters, the distribution of $\hat{\mathcal{C}}$ depends only on ρ and λ .

There is no known closed-form solution for the root of a general fifth order polynomial. I simulated from the asymptotic distribution for $\hat{\mathcal{C}}$ by the following method. Simulate a draw from the joint distribution of the five coefficients of the polynomial. Use a software package (I used Matlab version 5.3) to numerically calculate the roots of the resulting polynomial. The real root $\hat{\mathcal{C}}$ which maximizes $\mathcal{Q}^A(\mathcal{C}, \mathcal{B}(\mathcal{C}))$ is the simulated draw from the asymptotic distribution of $\hat{\mathcal{C}}$. The corresponding draw from the asymptotic distribution of the t-statistic is

$$\sigma_{\hat{\mathcal{C}}} \sqrt{\int W_c^2(t) dt + \frac{1}{2} \left(\mathcal{B}(\hat{\mathcal{C}}) \right)^2 - 2\mathcal{B}(\hat{\mathcal{C}}) \int tW_c(t) dt}.$$

The stochastic integrals were computed as the realizations of normalized sums of 500 successive draws from a discrete time Gaussian AR(1) process with autoregressive parameter $1 - c/500$.

The simulation procedure was repeated 100,000 times for each value of λ , ρ and c . The asymptotic critical value for a size $100\alpha\%$ test which rejects for small \hat{c} was calculated as the $100,000\alpha^{th}$ element of the vector of sorted draws for $\hat{\mathcal{C}}$. The power of the test at the alternative c was calculated as the proportion of draws below the critical value. A similar procedure was used to calculate the critical value and power of the test based on the t-statistic.

References

Chan, N. H. and Wei, C. Z. (1987), ‘Asymptotic inference for nearly nonstationary ar(1) processes’, *The Annals of Statistics* **15**(3), 1050–1063.

- Dufour, J. M. and King, M. L. (1991), ‘Optimal invariant tests for the autocorrelation coefficient in linear regressions with stationary or nonstationary ar(1) errors’, *Journal of Econometrics* **47**(1), 115–143.
- Elliott, G., Rothenberg, T. and Stock, J. H. (1996), ‘Efficient tests for an autoregressive unit root’, *Econometrica* **64**(4), 813–836.
- Elliott, G. and Stock, J. H. (2001), ‘Confidence intervals for autoregressive coefficients near one’, *Journal of Econometrics* **103**, 155–181.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986), *Robust statistics: the approach based on influence functions*, John Wiley and Sons, New York.
- Hansen, B. E. (1999), ‘The grid bootstrap and the autoregressive model’, *Review of Economics and Statistics* **81**, 594–607.
- Hasan, M. N. and Koenker, R. W. (1997), ‘Robust rank tests of the unit root hypothesis’, *Econometrica* **65**(1), 133–161.
- Herce, M. A. (1996), ‘Asymptotic theory of lad estimation in a unit root process with finite variance errors’, *Econometric Theory* **12**, 129–153.
- Hjort, H. L. and Pollard, D. (1993), ‘Asymptotics for minimizers of convex processes’, *Yale University Department of Statistics Preprint*.
- Hoek, H., Lucas, A. and van Dijk, H. K. (1995), ‘Classical and bayesian aspects of robust unit root inference’, *Journal of Econometrics* **69**, 27–59.
- Judd, K. L. (1998), *Numerical Methods in Economics*, The MIT Press, Cambridge.
- Lehmann, E. L. (1959), *Testing Statistical Hypotheses*, Wiley, New York.
- Lucas, A. (1995), ‘Unit root tests based on m estimators’, *Econometric Theory* **11**, 331–346.
- Phillips, P. C. B. (1987), ‘Toward a unified asymptotic theory for autoregression’, *Biometrika* **74**, 535–547.
- Phillips, P. C. B. (1988), ‘Regression theory for near-integrated time series’, *Econometrica* **56**, 1021–1043.
- Pollard, D. (1985), ‘New ways to prove central limit theorems’, *Econometric Theory* **1**, 295–314.
- Pollard, D. (1991), ‘Asymptotics for least absolute deviation regression estimators’, *Econometric Theory* **7**, 186–199.
- Rothenberg, T. J. (1984), Approximating the distributions of econometric estimators and test statistics, in Z. Griliches and M. Intriligator, eds, ‘Handbook of Econometrics’, Vol. 2, North Holland, Amsterdam.

- Rothenberg, T. J. and Stock, J. H. (1997), ‘Inference in a nearly integrated autoregressive model with nonnormal innovations’, *Journal of Econometrics* **80**, 269–286.
- Shin, D. W. and So, B. S. (1999), ‘Unit root tests based on adaptive maximum likelihood estimation’, *Econometric Theory* **15**, 1–23.
- Stock, J. H. (1991), ‘Confidence intervals for the largest autoregressive root in u.s. macroeconomic time series’, *Journal of Monetary Economics* **28**, 435–459.
- Thompson, S. B. (2001*a*), ‘Robust confidence intervals for autoregressive coefficients near one’, *Working Paper* .
- Thompson, S. B. (2001*b*), ‘Robust unit root testing with correct size and good power’, *Working Paper* .
- van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.
- Xiao, Z. (2001), ‘Likelihood-based inference in trending time series with a root near unity’, *Econometric Theory* **17**, 1082–1112.