EFWW
7/12/07
9:45 AM

Low-Frequency Robust Cointegration Testing*

Ulrich K. Müller
Princeton University
Department of Economics
Princeton, NJ, 08544

Mark W. Watson
Princeton University
Department of Economics
and Woodrow Wilson School
Princeton, NJ, 08544

Preliminary: June 2007

Abstract

Standard inference in cointegrating models is fragile for two distinct reasons. First, even though cointegration concerns low-frequency variability, inference relies on higher frequency variability in the data; second, inference assumes an $I(1)$ model for the common trends which may not accurately describe the data's persistence. This paper discusses efficient inference about the cointegrating vector in a bivariate model that is robust to both sources of misspecification. A small of number of weighted averages are used to summarize the data's low-frequency variability. These weighted averages have an asymptotic multivariate normal distribution, and cointegration imposes restrictions on the associated covariance matrix. Under the null hypothesis, the common trend is modeled by a flexible limiting Gaussian process that includes the $I(1)$, local-to-unity, and fractional model as special cases. The flexibility in the trend process introduces a large number of nuisance parameters. An upper bound on power for tests is presented, and this upper bound is computed for tests that control size for the flexible trend specification, and several special cases. A simple test motivated by the analysis in Wright (2000) almost achieves the power bound under the flexible common trend specification, so the test is approximately optimal.

JEL classification: C22, E32

Keywords: least favorable distribution, power envelope, nuisance parameters

*Support was provided by the National Science Foundation through grants SES-0518036 and SES-0617811.

1 Introduction

Many economic time series are individually highly persistent, but some linear combination might be much less persistent—this is the fundamental insight of cointegration. Accordingly, a suite of practical methods have been developed for conducting inference about cointegrating vectors, the coefficients that lead to this reduction in persistence. In their standard form, these methods assume that the persistence is the result of a common $I(1)$ stochastic trend, and that the error correction term, the non-persistent linear combination of the variables, is $I(0)$.[1] This $I(1)/I(0)$ dichotomy drives standard cointegration analysis, but may lead to fragile inference for two distinct reasons. First, the persistence reduction associated with moving from an $I(1)$ to an $I(0)$ process might be implausible in many applications. Second, standard methods rely critically on particular properties of the $I(1)$ process about which there may be uncertainty that cannot be resolved by examination of the data. This paper studies efficient inference methods for the cointegrating vector in a bivariate framework that address both these complications.

Consider first the issue that in the standard asymptotic reasoning, the error correction term and the stochastic trend are of different orders of persistence: apart from a scaling factor, the asymptotic behavior of $I(0)$ processes is no different from i.i.d. random variables in the sense that both satisfy a functional central limit theorem, while $I(1)$ processes are just like random walks in this sense. In practice, the dividing line between an persistent and non-persistent process is far less clear. Because cointegration is inherently about the low-frequency behavior of time series, persistence and non-persistence might more usefully be defined in terms of low frequency variability. Of course, this in turn requires a dividing line to define "low-frequencies", but natural definitions typically follow from the phenomenon under study. For example, in macroeconomics, long-run or low-frequency variability typically refers to frequencies lower than the business cycle, which are reasonably characterized by periodicities greater than 8 years. Thus, a macroeconomic time series might usefully be defined as $I(0)$ or "non-persistent" if it behaves like an i.i.d. process over frequencies with periods longer than 8 years, and otherwise it is "persistent". Müller and Watson (2006) use this idea to study univariate properties of economic time series, but the reasoning is equally

[1] See, for instance, Johansen (1988), Phillips and Hansen (1990), Saikkonen (1991), Park (1992) and Stock and Watson (1993).

(or more) compelling for cointegration.

As shown by Müller and Watson (2006), low-frequency variability can be summarized by a small number of weighted averages of the data, where the weights are low frequency trigonometric series. For example, only $q = 12$ weighted averages are needed to capture variability lower than the business cycle for time series that span 50 years (postwar data) regardless of the sampling frequency (months, quarters, weeks, etc.). Section 2 thus considers the behavior of these weighted averages as the sample size T grows large, but with q held fixed.[2] As in Bierens (1997), the weighted averages have a multivariate normal limiting distribution, and cointegration imposes restrictions on the covariance matrix of this distribution. Asymptotically, inference about the cointegrating vector thus becomes inference about the covariance matrix of a multivariate normal random vector.

An alternative to this low-frequency transformation approach is to model the persistence in the error correction term directly. A well developed body of work has pursued this approach in the fractional integration framework, where the error correction term is allowed to have long memory.[3] See, for instance, Jeganathan (1999), Kim and Phillips (2000), Robinson and Hualde (2003), Robinson and Marinucci (2003) and Velasco (2003). We believe the low-frequency transformation approach to be attractive beyond its statistical convenience because it explicitly acknowledges the relative scarcity of low-frequency information, it is robust to dynamic properties beyond the chosen frequency band, it does not require hard-to-interpret bandwidth choices, it is stable under aggregation, and it arguably gives the concept of "persistence" of an economic time series a straightforward interpretation.

The second important issue in cointegration analysis involves the uncertain nature of the common stochastic trend. Elliott (1998) provided a dramatic demonstration of the fragility of standard cointegration methods by demonstrating that they fail to control size when the common stochastic trend is not $I(1)$, but rather is "local-to-unity" in the sense of Bobkoski (1983), Cavanagh (1985), Chan and Wei (1987) and Phillips (1987).[4] The development of valid tests for a local-to-unity stochastic trend is complicated by the fact that the local-to-unity nuisance parameter cannot be consistently estimated. Cavanagh, Elliott, and Stock

[2] As discussed in further detail in Müller and Watson (2006), also Bierens (1997), Phillips (1998) and Müller (2006) consider time series inference based on a finite number of weighted averages.

[3] Canjels (1997) investigates the same idea with a local-to-unity specification of the error correction term.

[4] Also see Elliott and Stock (1994) and Jeganathan (1997).

(1995) propose several procedures to adjust critical values from standard tests to control size over a range of values of the local-to-unity parameter, and their general approach has been used by several other researchers; Campbell and Yogo (2006) provides a recent example. Stock and Watson (1996) and Jansson and Moreira (2006) go further and develop inference procedures with specific optimality properties in the local-to-unity model. In the fractional cointegration literature, the common stochastic trend is modelled as fractionally integrated, although the problem is different from the local-to-unity case as the fractional parameter can be consistently estimated under standard asymptotics. Yet, Müller and Watson (2006) demonstrate that at least based on below business cycle variation in the data, it is a hopeless endeavor to try to consistently discriminate between, say, local-to-unity and fractionally integrated stochastic trends.

As demonstrated by Wright (2000), it is nevertheless possible to conduct inference about the cointegrating vector without knowledge about the precise nature of the common stochastic trend. Wright's idea is to use the $I(0)$ property of the error correction term as the identifying property of the true cointegrating vector, so that inference about the cointegrating vector can be conducted using a stationarity test of the model's putative error correction term. Because the non-$I(0)$ data are not used in the analysis, Wright's procedures are robust in the sense that they control size under any model for the non-$I(0)$ data. But, it is unknown under what circumstances, if any, Wright's procedures are efficient.

Section 2 of this paper provides a formulation of the bivariate cointegrated model in which the common stochastic trend follows a flexible limiting Gaussian process that includes the $I(1)$, local-to-unity, and fractional/long-memory models as special cases. The model also allows more general time lags between the $I(0)$ component and stochastic trend than standard formulations of the cointegrated model. Throughout the paper, inference procedures are studied in the context of this general formulation of the cointegrated model. This may be viewed as a response to Granger's (1993) call to think of the persistence of macro time series as the result of a wide range of possible data generating processes beyond the $I(1)$ model, and to abandon attempts to identify the exact nature of the persistence process from the data.

The price to pay for this generality is that it introduces a potentially large number of nuisance parameters that complicate the derivation of efficient inference procedures. The nuisance parameters characterize the properties of the stochastic trend, the relationship be-

tween the stochastic trend and the model's $I(0)$ component, and the mapping from the stochastic trend and $I(0)$ component to the data. None of these nuisance parameters can be consistently estimated in the low-frequency framework introduced above because their only effect is on the covariance matrix of the limiting distribution of the weighted averages. Invariance considerations discussed in Section 3 makes a subset of these parameters irrelevant for the testing problem, but potentially many nuisance parameters remain. The main problem of this paper is thus to deal with this difficulty, and Sections 4 and 6 take up this issue.

Section 4 presents a general result concerning an upper bound on the power of tests when the null hypothesis involves a vector of nuisance parameters. This result, together with a numerical analysis described in Section 6, makes it possible to compute low upper power bounds (approximate "least upper power bounds") for tests concerning the values of cointegrating vectors. These bounds are computed for an alternative with the standard $I(1)$ stochastic trend, but under the constraint that the tests control size over a wide range of stochastic trend processes, ranging from the standard $I(1)$ model to a highly flexible model.

These power bounds are useful for at least two purposes. First, differences in the power bounds (interpreted as differences in least upper bounds) associated with restrictions on the trend process (for example, restricting the general stochastic trend process to be $I(1)$) quantify the information in the restriction about the value of the cointegrating vector. Second, and most importantly, they provide a bound on the power envelope for any asymptotically valid test. In particular, the bounds allow us to assess the relative efficiency of a low-frequency version of Wright's (2000) test that is introduced in Section 5. As it turns out, the power of this test essentially coincides with the power bound for the highly flexible version of the common trend process, and is close to the bound for several restricted, but still flexible common trend processes. Thus Wright's test—that is ignoring the non-$I(0)$ data—yields an essentially efficient test in absence of strong *a priori* knowledge about the nature of the persistence.

2 Model

2.1 Time Domain Representation of the Model

Consider a bivariate model where p_t, $t = 1, ..., T$, denotes the 2×1 vector of variables under study. This section begins with a time domain representation of the cointegrated model that differs from the standard model by allowing a flexible process for the common trend and flexible time lags in the relation between the trend and the model's $I(0)$ component. The low-frequency components of the model, which form the basis for inference, are discussed in the next subsection.

It is convenient to transform p_t so that one of its elements is $I(0)$ under the null hypothesis. Thus, let β denote the cointegrating vector with value β_0 under the null, and let $y_t = \beta_0' p_t$ denote the linear combination of the variables that is $I(0)$ under the null hypothesis, that is, y_t is the null model's error correction term. Let $x_t = \delta' p_t$ where δ is linearly independent of β_0, so that x_t is not $I(0)$ under the null. The variables y_t and x_t are the transformations of p_t that will be used in the analysis. While y_t is determined (up to scale) by the null hypothesis, x_t is not: different values of δ yield different definitions of x_t. Typical applications of cointegration do not lead to a natural definition of x_t, so that most extant inference procedures are invariant (or asymptotically invariant) to the transformations induced by different values of δ. Restrictions associated with invariant tests are discussed in detail in the next section, but for the purposes of the analysis in this section, assume that y_t and x_t are known linear combinations of the data with the property that y_t is $I(0)$ under the null.

It is convenient to represent (y_t, x_t) in terms of a common stochastic trend v_t, and an $I(0)$ variable z_t:

$$\begin{aligned} y_t &= \lambda_{yv} v_t + \lambda_{yz} z_t \\ x_t &= \lambda_{xv} v_t + \lambda_{xz} z_t \end{aligned} \tag{1}$$

In this representation, the restriction that y_t is $I(0)$ corresponds to the restriction $\lambda_{yv} = 0$. All of the test statistics discussed in this paper are invariant to adding constants to the observations, so that constant terms are suppressed. We also abstract from linear deterministic trends, but note that these could be incorporated in the model using modifications like those used in Müller and Watson (2006).

In the usual cointegration framework, the common stochastic trend v_t is an $I(1)$ process in the sense that $T^{-1/2}v_{[sT]} \Rightarrow \int_0^s dW_v(r)$, where '$\Rightarrow$' means weak convergence, W_v is a standard Wiener process, here and below the convergence is with respect to the Skorohod metric on the space of cadlag functions on the unit interval, and all limits are taken as $T \to \infty$. Because the scales of v_t and z_t are not separately identified from the λ coefficients in (1), the limits of v_t and z_t are represented in terms of scale normalized random processes. Local-to-unity generalizations such as those considered in Elliott (1998) assume that $T^{-1/2}v_{[sT]} \Rightarrow \int_0^s e^{-c(s-r)}dW_v(r)$.[5] We will use a similar, but more general representation of the process for or v_t. Specifically, we assume the asymptotic representation

$$T^{-\delta_v}v_{[sT]} \Rightarrow \int_{-\infty}^{s} g(s,r)dW_v(r) \tag{2}$$

where $T^{-\delta_v}$ is a scale factor and $g(s,r)$ is sufficiently well behaved to ensure that there exists a cadlag version of the process $\int_{-\infty}^{s} g(s,r)dW_v(r)$.[6] Special cases of this general specification are the $I(1)$ model (which correspond to $g(s,r) = \mathbf{1}[r \geq 0]$ and $\delta_v = 1/2$), the stationary local-to-unity model (which correspond to $g(s,r) = e^{-c(s-r)}$, $c > 0$ and $\delta_v = 1/2$), the type-I fractional model with parameter d (where $\delta_v = d + 1/2$, $g(s,r) = C(d)((s-r)^d - (-r)^d)$ for $r \leq 0$, $g(s,r) = C(d)(s-r)^d$ for $r > 0$ and $C(d) = \left(\frac{1}{2d+1} + \int_0^\infty \left[(1+\lambda)^d - \lambda^d\right]^2 d\lambda\right)^{-1/2}$, see Marinucci and Robinson (1999)), and the general stationary model (which corresponds to the restriction that $g(s,r) = \widetilde{g}(s-r)$ for some function $\widetilde{g}$). The generalization (2) produces processes with flexible low-frequency covariance properties and allows processes that are more or less persistent than the $I(1)$ process, but in less restricted ways than the local-to-unity or fractional models.

We assume that z_t is $I(0)$ in the sense that its partial sums satisfy

$$T^{-1/2}\sum_{t=1}^{[sT]} z_t \Rightarrow W_z(s) = \rho W_v(s) + (1-\rho^2)^{1/2}W_\varepsilon(s) \tag{3}$$

where W_ε is a standard Wiener process that is independent of $W_v(s)$ and $-1 \leq \rho \leq 1$ governs the long-run correlation between the innovations in the trend v_t and the $I(0)$ term z_t.

[5]This convergence requires v_t to be modelled as a double array process, but we do not indicate this explicity here and below to ease notation.

[6]For the local-to-unity process, this is a well known result, and a straightforward application of Kolmogorov's Continuity Theorem shows that this is the case for the processes considered in Sections 5 and 6 below.

Time lags in the relation between the error correction term and the common stochastic trend

With the exception of the more general stochastic trend process, (1), (2), and (3) are a standard representation of a cointegrated system. We digress here to offer an additional generalization of the standard representation.

The algebraic motivation for the generalization is simple enough. In (3), notice that the index for W_v is s and not, for example, $s+\phi$. The standard asymptotic representation thus imposes a negligible time lag (as a fraction of the sample size) between innovations in the stochastic trend and the $I(0)$ component. Yet, to take a concrete example, one can imagine macroeconomic processes in which a recession (interpreted as an $I(0)$ variation in the data) is associated with trend variation many years in the future, caused for example by human capital accumulation or other endogenous growth forces. This suggests that it is useful to allow for more general time lags between z_t and v_t.

There are several ways to accomplish this. One flexible way represents $W_v(s)$ as

$$W_v(s) = \sum_{i=1}^{n_\kappa} \kappa_i W_{v,i}(s) \tag{4}$$

where $W_{v,i}$ are independent standard Wiener processes and where κ_i are constants that satisfy $\sum_{i=1}^{n_\kappa} \kappa_i^2 = 1$. The process for v_t remains as (2), while the process for the partial sums of z_t are replaced with

$$T^{-1/2} \sum_{t=1}^{[sT]} z_t \Rightarrow W_z(s) = \rho \sum_{i=1}^{n_\kappa} \kappa_i W_{v,i}(s+\phi_i) + (1-\rho^2)^{1/2} W_\varepsilon(s) \tag{5}$$

where ϕ_i are non-zero constants. The limiting univariate stochastic processes for v_t and z_t are as before, but different choices for the parameters n_κ, κ_i, and ϕ_i produce flexible time lags in their cross covariances while preserving joint normality.

2.2 Low-Frequency Representation of the Model

Cointegration is a restriction on the low-frequency behavior of time series, and as discussed in the introduction, we therefore focus on the low-frequency behavior of (y_t, x_t). This low-frequency variability is summarized by a small number, q, of weighted averages of the data,

where the weights are low-frequency trigonometric series. Specifically, we use weights associated with the cosine transformation, where the weights associated with the $j'th$ weighted average are $\Psi_j(s) = \sqrt{2}\cos(j\pi s)$. For any sequence $\{a_t\}_{t=1}^T$, the $j'th$ weighted average will be denoted by

$$A_{Tj} = \int_0^1 \Psi_j(s) a_{[sT]+1} ds = \iota_{jT} T^{-1} \sum_{t=1}^{T} \Psi_j(\tfrac{t-1/2}{T}) a_t \tag{6}$$

where $\iota_{jT} = \frac{2T}{j\pi}\sin(\frac{j\pi}{2T}) \to 1$ for all fixed j.

As demonstrated by Müller and Watson (2006), the weighted averages A_{Tj}, $j = 0, \cdots, q$, essentially capture the variability in the sequence corresponding to frequencies below $q\pi/T$. Using this notation, the vectors $X_T = (X_{T1}, \ldots, X_{Tq})'$ and $Y_T = (Y_{T1}, \ldots, Y_{Tq})'$ summarize the variability in the data corresponding to frequencies lower than $q\pi/T$, where the component corresponding to $j = 0$ is excluded to make the results invariant to adding constants to the data. With $q = 12$, (Y_T, X_T) capture variability lower than the business cycle (periodicities greater than 8 years) for time series that span 50 years (postwar data) regardless of the sampling frequency (months, quarters, weeks, etc.) This motivates us to consider the behavior of these vectors as $T \to \infty$, but with q held fixed.

The large sample behavior of X_{Tj} and Y_{Tj} follows from the behavior of Z_{Tj} and V_{Tj}. Using the assumed limits (2) and (3), the definition of the cosine transformation (6), the continuous mapping theorem, and integration by parts for the terms involving Z_{Tj}, one obtains

$$\begin{bmatrix} T^{1/2}(Z_{T1}, \ldots, Z_{Tq})' \\ T^{-\delta_v - 1}(V_{T1}, \ldots, V_{Tq})' \end{bmatrix} \Rightarrow \begin{bmatrix} Z \\ V \end{bmatrix} = \begin{bmatrix} (Z_1, \ldots, Z_q)' \\ (V_1 \ldots, V_q)' \end{bmatrix} \tag{7}$$

where

$$Z_j = \int_0^1 \Psi_j(\lambda) dW_z(\lambda) \text{ and } V_j = \int_{-\infty}^1 \left(\int_{r \vee 0}^1 \Psi_l(\lambda) g(\lambda, r) d\lambda \right) dW_v(r).$$

so that

$$\begin{aligned} E[Z_l Z_j] &= \int_0^1 \Psi_l(\lambda)\Psi_j(\lambda) d\lambda = 1 \\ E[Z_l V_j] &= \rho \sum_{i=1}^{n_\kappa} \kappa_i^2 \int_{-\infty}^1 \Psi_l(r - \phi_i) \left(\int_{r \vee 0}^1 \Psi_j(\lambda) g(\lambda, r) d\lambda \right) dr \qquad (8) \\ E[V_l V_j] &= \int_{-\infty}^1 \left(\int_{r \vee 0}^1 \Psi_l(\lambda) g(\lambda, r) d\lambda \right) \left(\int_{r \vee 0}^1 \Psi_j(\lambda) g(\lambda, r) d\lambda \right) dr \qquad (9) \end{aligned}$$

for $l, j = 1, \cdots, q$, and thus

$$\begin{bmatrix} Z \\ V \end{bmatrix} \sim \mathcal{N}(0, \Sigma_{(Z,V)}) \quad \text{where } \Sigma_{(Z,V)} = \begin{bmatrix} I_q & \Sigma_{zv} \\ \Sigma_{vz} & \Sigma_{vv} \end{bmatrix}.$$

Defining $l_{yv} = T^{-\delta_v - 1/2}\lambda_{yv}$, $l_{yz} = \lambda_{yz}$, $l_{xv} = \lambda_{xv}$ and $l_{xz} = T^{-\delta_v - 1/2}\lambda_{xz}$, with $(l_{yv}, l_{yz}, l_{xv}, l_{xz})$ held fixed as $T \to \infty$,

$$\begin{bmatrix} T^{1/2}Y_T \\ T^{-\delta_v - 1}X_T \end{bmatrix} \Rightarrow \begin{bmatrix} (Y_1, \ldots, Y_q)' \\ (X_1, \ldots, X_q)' \end{bmatrix} = \begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} l_{yv}V + l_{yz}Z \\ l_{xv}V + l_{xz}Z \end{bmatrix} \tag{10}$$

follows directly from (1) and (7). Invariance considerations discussed in the next section make it useful to reparametrize the model in terms of the scalar b and $\gamma = (\gamma_1, \gamma_2, \gamma_3)$, where $b = l_{yv}/l_{yz}$, $\gamma_1 = l_{yz}$, $\gamma_2 = l_{xz}/l_{yz}$, and $\gamma_3 = l_{xv} - l_{xz}l_{yv}/l_{yz}$, so that

$$\begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} \gamma_1 I_q & 0 \\ \gamma_1\gamma_2 I_q & \gamma_3 I_q \end{bmatrix} \begin{bmatrix} Z + bV \\ V \end{bmatrix} \sim N(0, \Sigma_{(Y,X)}). \tag{11}$$

The null hypothesis that y_t is $I(0)$ implies that $b = 0$, but imposes no restrictions on $\gamma_1, \gamma_3 \in \mathbb{R}\backslash\{0\}$ and $\gamma_2 \in \mathbb{R}$. Using this notation, $\Sigma_{(Y,X)}$ can be written as

$$\Sigma_{(Y,X)} = \begin{bmatrix} \gamma_1 I_q & 0 \\ \gamma_1\gamma_2 I_q & \gamma_3 I_q \end{bmatrix} \Sigma(b) \begin{bmatrix} \gamma_1 I_q & 0 \\ \gamma_1\gamma_2 I_q & \gamma_3 I_q \end{bmatrix}' \tag{12}$$

$$\text{where } \Sigma(b) = \begin{bmatrix} I_q + b(\Sigma_{zv} + \Sigma_{vz}) + b^2\Sigma_{vv} & \Sigma_{zv} + b\Sigma_{vv} \\ \Sigma_{vz} + b\Sigma_{vv} & \Sigma_{vv} \end{bmatrix}.$$

The covariance matrix $\Sigma_{(Y,X)}$ depends on b, with $b = 0$ under the null, the three dimensional parameter γ which determines the mapping from the canonical $I(0)$ and trend variables (Z, V) to the observations (Y, X), and a parameter, say, θ that determines Σ_{vv} and Σ_{vz} via (8) and (9). The parameter vector θ includes $(\rho, \{\kappa_i, \phi_i\})$ and any additional parameters that characterize the function $g(s, r)$. The hypothesis testing problem $H_0 : b = 0$ corresponds to a restriction on the covariance matrix of a multivariate normal distribution, where the covariance matrix also depends on a set of nuisance parameters (γ, θ). The nuisance parameters complicate the testing problem and are discussed in the following two sections. In the next section, we argue that invariance considerations lead naturally to tests that do not depend on γ. Section 4 takes up the problem associated with the nuisance

parameters θ. Before launching into that discussion, it is useful get some insight into how θ affects the structure of $\Sigma_{(Y,X)}$ by considering two extreme cases.

The first case is the standard cointegration model: $g(s,r) = \mathbf{1}[r \geq 0]$, so that the trend follows an $I(1)$ process, and there are no time lags (so that z_t is as in (3)). In this case Σ_{vv} is a diagonal matrix with ith diagonal element proportional to i^{-2}, so that the elements of V are independent but severely heteroskedastic with lower frequency components having higher variances (reflecting the familiar spectral analysis intuition that autocorrelation is transformed into heteroskedasticity). The i,jth element of Σ_{vz} is equal to 0 when i and j are both even or odd, and proportional to $\rho/(i^2-j^2)$ otherwise, so that many of the elements of V and Z are independent. In this formulation, the covariance matrix is completely determined by the parameter ρ, so that θ contains only one element.

In the second case, suppose $g(s,r)$ is unrestricted, so a highly flexible model for the trend is allowed, although again without time lags. The following lemma shows that when there are no *a priori* restrictions on $g(s,r)$, the only restriction on $\Sigma_{(V,Z)}$ is that $\Sigma_{zz} = I_q$.

Lemma 1 *For any $2q \times 2q$ positive definite matrix Σ^* with upper left $q \times q$ block equal to I_q there exists a version of model (2) and (3) with $\rho = 1$ such that $\Sigma_{(Z,V)} = \Sigma^*$.*

Proof. *See Appendix.* ■

Lemma 1 implies that when there are no *a priori* restrictions on $g(s,r)$ there is no loss in generality in choosing θ as the $q^2 + q(q+1)/2$ parameters that directly determine Σ_{vv} and Σ_{zv}.

3 Invariance

Because $\int_0^1 \Psi_j(r)dr = 0$ for $j = 1, ..., q$, the low-frequency transformations (Y_T, X_T) are invariant to transformations $(y_t, x_t) \rightarrow (y_t + c_y, x_t + c_x)$ for arbitrary constants c_x and c_y. This section considers invariance to the scale of the variables (Y_T, X_T) and to the transformations of the original data p_t that determine x_t. More precisely, we consider tests that are invariant to the group of transformations

$$\begin{bmatrix} Y_T \\ X_T \end{bmatrix} \rightarrow \begin{bmatrix} a_{yy} Y_T \\ a_{xy} Y_T + a_{xx} X_T \end{bmatrix} \tag{13}$$

where a_{yy} and a_{xx} are non-zero. This is equivalent to transforming (y_t, x_t) in an analogous fashion, that is, it allows arbitrary linear transformations of (y_t, x_t) that maintain the restriction that the first element is $I(0)$ under the null hypothesis. It is straightforward to check that the $(2q-3)\times 1$ vector

$$Q_T = (\frac{Y_{T,1}}{Y_{T,q}}, \cdots, \frac{Y_{T,q-1}}{Y_{T,q}}, \frac{X_{T,1} - \frac{X_{T,q}}{Y_{T,q}}Y_{T,1}}{X_{T,q-1} - \frac{X_{T,q}}{Y_{T,q}}Y_{T,q-1}}, \cdots, \frac{X_{T,q-2} - \frac{X_{T,q}}{Y_{T,q}}Y_{T,q-2}}{X_{T,q-1} - \frac{X_{T,q}}{Y_{T,q}}Y_{T,q-1}})' \tag{14}$$

is a maximal invariant of (13), and by (10) and the continuous mapping theorem,

$$Q_T \Rightarrow Q = (\frac{Y_1}{Y_q}, \cdots, \frac{Y_{q-1}}{Y_q}, \frac{X_1 - \frac{X_q}{Y_q}Y_1}{X_{q-1} - \frac{X_q}{Y_q}X_q}, \cdots, \frac{X_{q-2} - \frac{X_q}{Y_q}Y_{q-2}}{X_{q-1} - \frac{X_q}{Y_q}Y_{q-1}})'. \tag{15}$$

By Theorem 6.2.1 of Lehmann and Romano (2005), any test that is invariant to (13) can be written as a function of the maximal invariant Q_T. The following Lemma provides an expression for the density of the limiting random vector Q.

Lemma 2 *Let $f_\theta(\cdot,\cdot)$ be the density of $(Z + bV, V)$, and suppose $\Sigma(b)$ is full rank. The probability density of Q is equal to*

$$\begin{aligned} & C\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\gamma_1\gamma_3|^{-q+1} f_\theta(\gamma_1^{-1}Y, \gamma_3^{-1}(X - \gamma_2 Y))d\gamma_1 d\gamma_2 d\gamma_3 \\ = \;& C' A_Y^{-1/2}(Y'\Sigma(b)_{22}^{-}Y)^{q-3/2}\left[(\sqrt{A_X A_Y} - A_{XY})^{1-q} + (\sqrt{A_X A_Y} + A_{XY})^{1-q}\right] \end{aligned}$$

where the constants C and C' do not depend on (Y, X), $A_X = (X'\Sigma(b)_{22}^{-}X)(Y'\Sigma(b)_{22}^{-}Y) - (Y'\Sigma(b)_{22}^{-}X)^2$, $A_Y = (Y'\Sigma(b)_{11}^{-}Y)(Y'\Sigma(b)_{22}^{-}Y) - (Y'\Sigma(b)_{12}^{-}Y)^2$ and $A_{XY} = (Y'\Sigma(b)_{12}^{-}Y)(X'\Sigma(b)_{22}^{-}Y) - (Y'\Sigma(b)_{12}^{-}X)(Y'\Sigma(b)_{22}^{-}Y)$ with $\Sigma(b)_{ij}^{-}$ the j,lth $q\times q$ block of $\Sigma(b)^{-1}$ for $j,l = 1,2$.

Proof. *See Appendix.* ■

The lemma shows that the density of Q does not depend on γ, so that the restriction to invariant tests takes care of this nuisance parameter. Furthermore, Lemma 2 provides a closed-form formula for the density of Q as a function of Y and X, which simplifies the the construction and analysis of efficient tests in the following sections.

Returning to the testing problem, recall that covariance matrix $\Sigma_{(Y,X)}$ depends on three sets of parameters (b, γ, θ), where b is the parameter of interest (with $b = 0$ under the null) and (γ, θ) are nuisance parameters. This section has argued that invariance considerations make the nuisance parameters in γ irrelevant for the testing problem. The next section discusses complications associated with the nuisance parameters that make up θ.

4 Power Bounds

4.1 A general result about power bounds

The basic version of the hypothesis testing problem that we are facing is a familiar one: Let U denote a single observation of dimension $k \times 1$. (In our problem, Q will play the role of U.) Under the null hypothesis U has probability density $f_\theta(u)$ with respect to the measure μ, where θ is a vector of nuisance parameters.[7] (In our problem, $f_\theta(u)$ is the density of Q derived in Lemma 2 above, that depends on θ through the matrix $\Sigma(b)$, and μ is Lebesgue measure on $\mathbb{R}^{2q-3}$.) Under the alternative, U has known density $h(u)$. (Specific choices for $h(u)$ for our problem are discussed below.) Thus, the null and alternative hypothesis are

$$\begin{aligned} H_0 &: \text{The density of } U \text{ is } f_\theta(u),\ \theta \in \Theta \\ H_1 &: \text{The density of } U \text{ is } h(u), \end{aligned} \tag{16}$$

and (possibly randomized) tests are measurable functions $\varphi : \mathbb{R}^k \mapsto [0,1]$, where $\varphi(u)$ is the probability of rejecting the null hypothesis when observing $U = u$, so that size and power are given by $\sup_{\theta\in\Theta} \int \varphi(u) f_\theta(u) d\mu(u)$ and $\int \varphi(u) h(u) d\mu(u)$, respectively. The aim is to construct an efficient test φ^*. Unfortunately, there does not exist a general method to construct such an efficient test φ^* in the presence of the nuisance parameter θ.

We suggest a practical approach to this problem. We derive a set of upper bounds on the power of tests of (16), and then use numerical methods to choose a low upper bound from this set. The resulting upper bound may be used in two ways. First, when the dimension of θ is small, we develop an algorithm to numerically identify a test that has only marginally smaller power than the power bound. For practical purposes, this test is optimal. Second, we use the upper power bound to assess the efficiency of an *ad hoc* test that is known to control size over $\theta \in \Theta$. If the *ad hoc* test's power is close to the power bound, then again, one has identified an approximately optimal test.

A standard device for problems such as (16) is to consider a Neyman-Pearson test for a related problem in which the null hypothesis is replaced with a mixture

$$H_\Lambda : \text{The density of } U \text{ is } \int f_\theta(u) d\Lambda(\theta)$$

[7] And we assume $f_\theta(u)$ to be jointly Borel-measurable in θ and u.

where Λ is a probability distribution for θ. The following lemma shows that the Neyman-Pearson test for H_Λ versus H_1 provides an upper power bound for tests of H_0 versus H_1.

Lemma 3 *Let φ_Λ be the best level α test of H_Λ against H_1. Then for any level α test φ of H_0 against H_1,$\int \varphi_\Lambda(u)h(u)d\mu(u) \geq \int \varphi(u)h(u)d\mu(u)$.*

Proof. *Since φ is a level α test of H_0, $\int f_\theta(u)d\mu(u) \leq \alpha$ for all $\theta \in \Theta$. Therefore, $\int\int f_\theta(u)d\mu(u)d\Lambda(\theta) = \int\int f_\theta(u)d\Lambda(\theta)d\mu(u) \leq \alpha$ (where the change in the order of integration is allowed by Fubini's Theorem), so that φ is also a level α test of H_Λ against H_1. The result follows by the definition of a best test.* ■

This result is closely related to Theorem 3.8.1 of Lehmann and Romano (2005) that provides conditions under which a least upper bound on the power for tests H_0 versus H_1 is associated with a "least favorable distribution" for θ, and that using this distribution for Λ produces the least upper power bound. The least favorable distribution Λ^* has the characteristic that the resulting φ_{Λ^*} is a level α test for testing H_0 versus H_1. Said differently, if φ_{Λ^*} is the best level α test of H_{Λ^*} against H_1 *and* is also a level α test for testing H_0 versus H_1, then $\varphi^* = \varphi_{\Lambda^*}$, that is φ_{Λ^*} is the most powerful level α test of H_0 versus H_1. Unfortunately, while the test associated with the least favorable distribution solves the testing problem (16), there is no general and constructive method for finding the least favorable distribution Λ^* (and it does not always exist).

With this in mind, Lemma 3 is stated so that Λ is not necessarily the least favorable distribution. That is, the bound in Lemma 3 holds for any distribution Λ. The goal of the numerical analysis carried out below is to choose Λ to approximate the least upper bound. Importantly, though, even if one cannot identify the least favorable distribution, Lemma 3 shows that the power of φ_Λ provides a valid bound for the power of any test of H_0 versus H_1, for any Λ.

4.2 Bounds on the asymptotic weighted average power of invariant tests of cointegration

The preceding subsection has discussed a general upper bound for the exact small sample power of tests against a single alternative hypothesis. We now discuss in detail how this

result is useful for the problem of deriving bounds on the asymptotic power of low-frequency tests of cointegration, i.e. tests of the null hypothesis $H_0 : b = 0$.

As discussed in Section 4, the requirement that tests are invariant to the transformations (13) implies that they can be written as functions of the maximal invariant Q_T defined in (14). Let the (measurable) function $\varphi_T : \mathbb{R}^{2q-3} \mapsto [0, 1]$ be a test, where $\varphi_T(Q_T)$ denotes the rejection probability of the test given observations Q_T. Denote by $\mathrm{RP}_T(\varphi_T; M, b, \theta)$ the overall rejection probability of the test faced with data from model M with $\theta \in \Theta$ and $b \in \mathbb{R}$, that is $\mathrm{RP}_T(\varphi_T; M, b, \theta) = E\varphi_T(Q_T)$, where the expectation is with respect to the data generating process of model M with parameter values θ and b. As discussed in Section 2.2, possible models M are the classical $I(1)$ stochastic trend model, where $\theta = \rho$, the local-to-unity generalization with $\theta = (\rho, c)$, etc. Lemma 2 implies that the distribution of the limiting random vector Q depends on (b, θ), and so will the asymptotic power of invariant tests in general. Thus there does not exist an asymptotically uniformly most powerful test for the hypothesis testing problem

$$H_0 : b = 0 \quad \text{and } M = M_0,\ \theta \in \Theta_0 \quad \text{against} \quad H_1 : b \neq 0 \quad \text{and } M = M_1,\ \theta \in \Theta_1 \tag{17}$$

over (b, θ) in general.

We solve this difficulty by comparing the weighted average power of tests with some weighting function Γ over (b, θ). Formally, let

$$\mathrm{WAP}_T(\varphi_T) = \int \mathrm{RP}_T(\varphi_T; M_1, b, \theta) d\Gamma(b, \theta) \tag{18}$$

where Γ is an integrable weighting function on $\mathbb{R} \times \Theta_1$, and Θ_1 denotes the nuisance parameter space of θ in model M_1. The choice of M_1 and Γ govern for what kind of alternatives a test φ_T with large $\mathrm{WAP}_T(\varphi_T)$ is a good test, and we discuss specific choices below.

We defined our models in Section 2 in terms of their asymptotic properties; accordingly, we are interested in tests which control asymptotic size over $\theta \in \Theta_0$ for some model $M = M_0$, i.e. tests φ_T for which

$$\sup_{\theta \in \Theta_0} \limsup_{T \to \infty} \mathrm{RP}_T(\varphi_T; M_0, 0, \theta) \leq \alpha \quad \text{for all models satisfying (15).} \tag{19}$$

Tests that satisfy (19) are restricted to control asymptotic size for *all* data generating processes that lead to the weak limit (15), which is a potentially smaller set of tests than, say,

the class of tests that satisfy $\sup_{\theta\in\Theta_0} \limsup_{T\to\infty} \mathrm{RP}_T(\varphi_T; M_0, 0, \theta) \leq \alpha$ under additional assumptions on the distributions of v_t and z_t. The restriction to tests satisfying (19) is enough to yield a tight link between properties of tests for the hypothesis testing problem (17), which is inherently asymptotic in nature, and exact small sample hypothesis tests studied in Lemma 2 above.

Theorem 1 *Let the probability distributions $\Lambda : \Theta_0 \mapsto \mathbb{R}$ and $\Gamma : \mathbb{R} \times \Theta_1 \mapsto \mathbb{R}$ be such that $\Sigma(0)$ and $\Sigma(b)$ are full rank almost surely under Λ in model M_0, and under Γ in model M_1, respectively, and denote by $f_Q(M, b, \theta)$ the probability density of Q in model M and parameters (b, θ) defined in Lemma 2. Consider the exact small sample hypothesis problem*

$$\begin{aligned} &H_0 : Q \text{ has density } f_Q(M_0, 0, \theta) \quad \text{for } \theta \in \Theta_0 \\ &H_1 : Q \text{ has density } h \equiv \textstyle\int f_Q(M_1, b, \theta) d\Gamma(b, \theta). \end{aligned} \tag{20}$$

(i) Any almost everywhere continuous test φ of size α of (20) is an asymptotically valid test of (17), i.e. satisfies (19), and $\lim_{T\to\infty} WAP_T(\varphi) = \int \varphi h dQ$.

(ii) Let φ_Λ be the level α likelihood ratio test of $H_\Lambda : Q$ has density $\int f_Q(M_0, 0, \theta) d\Lambda(\theta)$ against H_1 in (20). Then for any test φ_T of (17) satisfying (19),

$$\limsup_{T\to\infty} WAP_T(\varphi_T) \leq \int \varphi_\Lambda h dQ.$$

Proof. See Appendix. ■

Theorem 1 is useful, as it shows that an understanding of the single null distribution/single alternative distribution hypothesis testing problem (20) concerning the limiting random vector Q has immediate implications for the hypothesis testing problem of interest (17): Part (i) implies that most size α tests of interest of (20) also control asymptotic size in (17), and have the same weighted average power. Part (ii) shows that power bounds constructed as in Lemma 3 for the hypothesis problem (20) also yield bounds on the asymptotic weighted average power in the class of tests of (17) which control size in the sense of (19). Tests for the problem (20) and their power are relatively easily computed, at least as long as Λ and Γ are discrete distributions, since by Lemma 2, the likelihood ratio statistic becomes the ratio of weighted averages of the quadratic forms in the jointly Gaussian vectors (Y, X).

We employ the result of Theorem 1 in two ways: First, write $\varphi_\Lambda = \mathbf{1}[\tau_\Lambda(Q) > k_\Lambda]$, where $\tau_\Lambda(Q)$ is the likelihood ratio statistic and k_Λ is the level α critical value for testing H_Λ

against H_1. Suppose that a marginal adjustment in the critical value leads to a level α test for (20), that is, suppose that $\varphi_\Lambda^c = \mathbf{1}[\tau_\Lambda(Q) > k_\Lambda + \varepsilon]$ is of size α in (20), where ε is a small number. If the power of φ_Λ is a continuous function of the critical value, then the power of φ_Λ^c will be only marginally lower than φ_Λ. If this is the case, then φ_Λ is approximately the least upper bound, and φ_Λ^c is an approximately optimal test for (20). By Theorem 1 (i), these properties of φ_Λ and φ_Λ^c immediately translate into the corresponding asymptotic size and weighted average power bound properties in the hypothesis test of interest (17). Pursuing this strategy requires verifying that φ_Λ^c has rejection probability of at most α for all $\theta \in \Theta_0$ in (20). This is a feasible numerical task when the dimension of θ is small, but is computationally intractable if the dimension of θ is large.

Second, suppose there exists a size α *ad hoc* test of (20). If one is able to identify a distribution Λ for which the power of φ_Λ is only marginally higher than the power of the *ad hoc* test for a weighted average power specification of interest, then the *ad hoc* test is essentially optimal, as no test can have substantially higher asymptotic weighted average power. In our problem, the low-frequency version of Wright's (2000) turns out to be such an *ad hoc* test.

5 An *ad hoc* similar invariant test

We study a version of the test suggested in Wright (2000). Wright was motivated by Elliott's (1998) critique of standard cointegration inference methods which showed that procedures that utilized an $I(1)$ model for the stochastic trend could have serious size distortions when the stochastic trend instead follows a local-to-unity process. Wright's proposal was to ignore the information in x_t, and investigate the null hypothesis by testing whether the putative error correction term y_t is $I(0)$. By ignoring x_t, Wright simplified the testing problem by eliminating the nuisance parameters θ, making Elliott's critique moot. While this resulted in obvious benefits, the costs—the resulting loss in power—are not obvious. These costs are quantified in the next section.

Wright's (2000) employs the locally optimal Nyblom (1989)/Kwiatkowski, Phillips, Schmidt, and Shin (1992) stationarity test against "local-level model" alternatives to test the $I(0)$ property of the putative error correction term. We implement Wright's suggestion using the low-frequency stationarity test developed in Müller and Watson (2006), which is the

low-frequency analogue of Elliott and Müller's (2006) point-optimal test in the local-level model. We will refer to this test as the JW ("Jonathan Wright") test. Specifically, the test is the efficient scale invariant test of

$$\begin{aligned} &H_0 : Y = \gamma_1 Z \sim N(0, \gamma_1^2 I_q) \quad \text{against} \\ &H_1 : Y = \gamma_1 (Z + \bar{b} V) \sim N(0, \gamma_1^2 (I_q + \bar{b}(\Sigma_{zv} + \Sigma_{vz}) + \bar{b}^2 \Sigma_{vv})) \end{aligned} \tag{21}$$

where V is the low frequency transformation of the $I(1)$ model without time shift, i.e. V is based on model (2) and (3) with $g(s,r) = \mathbf{1}[r \geq 0]$ and the implemented using $\bar{b} = 8$. A straightforward calculation using (8) shows that $\Sigma_{zv} + \Sigma_{vz} = 0$ for all ρ and q in this model, so that the resulting test is uniformly most powerful for (21) over ρ.

Following Müller and Watson (2006), the resulting test statistic, written as a function of observables Y_T, is

$$\text{JW} = \left(\sum_{l=1}^{q} Y_{T,l}^2 \right) / \left(\sum_{l=1}^{q} \frac{Y_{T,l}^2}{1 + \bar{b}^2/(\pi l)^2} \right) \tag{22}$$

and the null hypothesis is rejected for large values of JW. Table 1 provides asymptotically justified 1%, 5% and 10% critical values for $q = 6, \cdots, 18$. It is straightforward to see that the JW test is invariant to the group transformations (13) ($Y_T \to a_{yy} Y_T$ because of scale invariance and $X_T \to a_{xy} Y_T + a_{xx} X_T$ because the test statistic does not depend on X_T) and that the test is similar (the asymptotic distribution of the test statistic under the null hypothesis does not depend on θ for any null model M_0, because the test statistics does not depend on X_T). Thus the test will serve as an asymptotically valid invariant test of (17), and its asymptotic weighted average power obviously is a lower bound on the asymptotic weighted average power of any efficient test of (17). At the same time, the JW test is *ad hoc* because it ignores the potentially valuable information contained in X_T.

6 Practical issues

6.1 Specification of the alternative hypothesis

In the following computations, we limit the analysis to alternatives in which v_t is $I(1)$ without time shifts, so that the model M_1 in (17) is (2) and (3) with $g(s,r) = \mathbf{1}[r \geq 0]$. This is partly out of practical considerations: while there is a wide range of potentially interesting $g(s,r)$

and timing specifications, the computations for any particular specification are involved, and these computational complications limit the number of alternatives we can usefully consider. At the same time, one might consider the classical $I(1)$ model as an important benchmark against which it is interesting to maximize power—not necessarily because this is the only possible model under the alternative, but because a test that performs well against this alternative presumably has reasonable power properties for a range of empirically relevant models. It is important to stress that this specification of the alternative hypothesis has no bearing of the range of models that we allow under the null hypothesis, which we discuss below. We will compute power bounds on efficient tests of an $I(1)$ alternative that control size over a wide range of models of persistence, so our approach might be called *robust cointegration testing.*

Even with the alternative model of persistence specified, there remains the issue that asymptotic power of tests of (17) depends on the values of b and ρ. For these two parameters, we consider alternatives with $b = b_1$ and $\rho = \rho_1$ for a range of values for b_1 and ρ_1, so that Γ in (18) is degenerate with all mass at (b_1, ρ_1); the asymptotic power bounds derived via Theorem 1 then serve as bounds on the asymptotic power envelope over these values of b and ρ. Invariance reduces the dimension the problem somewhat, as the power of optimal tests may be seen from Lemma 2 to depend on b and ρ only through $|b|$, $|\rho|$, and the sign of ρb.[8]

6.2 Parameterization of $\Sigma_{(Y,X)}$ under the null hypothesis

The covariance matrix $\Sigma_{(Y,X)}$ under the null hypothesis depends on the model M_0 for the stochastic trend, such as the $I(1)$ model, the local-to-unity model, etc., and the parameters that characterize the common trend's weight function $g(s,r)$ and the parameters $(\rho, \{\kappa_i, \phi_i\}_{i=1}^{n_\phi})$ that govern the interaction between the $I(0)$ process and the trend. We consider seven models M_0 associated with different sets of parameters $\theta \in \Theta_0$. These seven models are associated with three specifications of $g(s,r)$ and two specifications of $\{\kappa_i, \phi_i\}_{i=1}^{n_\phi}$, which results in six specifications. The seventh specification leaves $g(s,r)$ unrestricted, which by Lemma 1 amounts to leaving Σ_{vz} and Σ_{vv} unrestricted.

[8]Invariance means that the test is invariant to, for example, multiplying v_t by -1 for all t. This changes the sign of b and ρ, but does not change the sign of $b\rho$.

The three restricted models for $g(s,r)$ are the $I(1)$, local-to-unity, and stationary models. The $I(1)$ model has $g(s,r) = \mathbf{1}[r \geq 0]$. The local-to-unity model has

$$g(s,r) = \begin{cases} e^{c(s-r)},\ c < 0 \\ \mathbf{1}[r \geq 0]e^{c(s-r)},\ c \geq 0 \end{cases}$$

where, as shown in Müller and Watson (2006), $\Sigma_{(Y,X)}$ is continuous at $c = 0$ despite the discontinuity in $g(s,r)$. The stationary model has $g(s,r) = \widetilde{g}(s-r)$, where the function $\widetilde{g}$ is parameterized as the step function

$$\widetilde{g}(x) = \sum_{i=1}^{n_{\widetilde{g}}} \widetilde{\theta}_i \mathbf{1}\left[\frac{i-1}{n_{\widetilde{g}}+1} \leq \frac{x}{1+x} < \frac{i}{n_{\widetilde{g}}+1}\right]. \tag{23}$$

The steps in $\widetilde{g}(x)$ occur at the points $\frac{i}{n_{\widetilde{g}}+1-i}$ so that more flexibility is allowed for small values of x (half of the points are associated with values of x less than 1, for example), and the specification (23) sets $\widetilde{g}(x) = 0$ for $x > n_{\widetilde{g}}$. In the numerical analysis, we choose $n_{\widetilde{g}} = 40$.

The time lags are chosen as $\phi = (\frac{-m+1}{m}, \frac{-m+2}{m}, ..., \frac{-1}{m}, 0, \frac{1}{m}, ...\frac{m-2}{m}, \frac{m-1}{m})'$ where $m = (n_\kappa - 1)/2$, which allows for a grid of time lags between -1 and 1. The values of κ_i are restricted only by $\Sigma_{i=1}^{n_\kappa} \kappa_i^2 = 1$. The numerical analysis uses $n_\kappa = 1$, which corresponds to the standard model with no time lags, and $n_\kappa = 39$, which allows considerable flexibility in the timing relationship between z_t and v_t.

The final parameter, ρ, is allowed to take on any value between 0 and 1, where the restriction to non-negative values is without loss of generality because of the invariance of the testing problem.

6.3 Computing upper power bounds

By Theorem 1, bounds on the power in the exact small sample hypothesis problem involving Q with an appropriate single alternative hypothesis also yield bounds on the asymptotic weighted average power of asymptotically valid invariant tests of (17). The classic results concerning the least favorable distribution for the nuisance parameter $\theta \in \Theta$ for the generic hypothesis problem (16) discussed in Section 4.1 then suggests the problem that we would ideally like to solve: find Λ, a probability distribution on Θ, such that the optimal level α test for H_Λ versus H_1 is also a level α test for H_0 versus H_1. The resulting test is an optimal test of H_0 versus H_1, and the resulting power provides a least upper power bound

for tests of H_0 versus H_1. Finding the distribution Λ that solves this problem is a formidable numerical task in problems such as ours. For example, for any particular value of θ, the covariance matrices Σ_{vz} and Σ_{vv} must be computed via (8) and (9), and the test's rejection frequency under the null and alternative must be computed using a numerical method such as Monte Carlo integration. Because the rejection frequency under the null hypothesis must be controlled uniformly over $\theta \in \Theta_0$, this is a daunting computational task. Using an algorithm described in the appendix (which is related to the one developed in Srianananthakumar and King (2006)), we compute an approximate solution to this problem when the dimension of θ is small. However, the dimension of θ can be very large in our problem—with $q = 12$, the model with $g(s, r)$ unconstrained leads to θ of dimension 222—and finding the least upper power bound becomes a numerically intractable problem.

This motivates us to suggest a computationally practical method for computing a *low* (as oppose to *least)* upper power bound. The method restricts Λ so that it has non-zero mass on a single point, say θ^* which is chosen so that the null distribution of (Y, X) is close to the distribution under the alternative. To be specific, let Σ_1 denote the covariance matrix of (Y, X) from (12) under a specific $I(1)$ alternative as described in Section 5.1 above (that is, for specific values of $b = b_1$ and $\rho = \rho_1$, and, say, $\gamma_1 = \gamma_3 = 1$ and $\gamma_2 = 0$), and let $\Sigma_0(\theta, \gamma)$ be the covariance matrix of (Y, X) under model M_0 of the null hypothesis, which depends on $\gamma \in \mathbb{R}^3$ and $\theta \in \Theta_0$. Denote the Kulback-Leibler divergence between the $2q \times 1$ distributions $N(0, \Sigma_1)$ and $N(0, \Sigma_0)$ as $K[\Sigma_1, \Sigma_0] = \frac{1}{2}(\ln(|\Sigma_1|/|\Sigma_0|) + \text{tr}(\Sigma_0^{-1}\Sigma_1) - 2q)$. The value of θ^* is chosen to numerically solve

$$\min_{\theta \in \Theta_0, \gamma \in \mathbb{R}^3} K[\Sigma_1, \Sigma_0(\theta, \gamma)], \tag{24}$$

that is, θ^* numerically minimizes the Kullback-Leibler divergence (or KLIC) between the distribution of specific alternative distribution (Y, X) and the null density as a function of θ and γ. While the minimization problem is over a large number of parameters—recall that θ contains as many as 222 parameters—the objective function is quickly computed and well behaved, so that numerical minimization is feasible.[9]

While this power is not a least upper power bound of tests of H_0 versus H_1, Lemma 3

[9]And as discussed in Section 4, the validity of the power bound from Theorem 1 based on the (degenerate) distribution Λ that puts all mass on the numerical minimizer in no way depends on a claim that is is the actual global minimizer.

implies that it produces an upper power bound, and the numerical results in the next section suggest that it is close the least upper power bound in the models where the least upper bound can be directly approximated. In many other cases it is close to the power of the JW test, so again it is close the least upper power bound (as the least upper bound cannot be below the power of the JW test). Thus, the power bound associated with this restricted problem proves to be a practical and useful approximation to the least upper power bound in this problem.

7 Results

Table 2 summarizes the power bounds for 5% level tests with $q = 12$, where panel A shows results for the alternative with $|b_1| = 7$, panel B shows results for $|b_1| = 14$, and each panel includes results for $|\rho_1| = 0.0$, 0.5, and 0.9. The first set of entries in each table correspond to power bounds constructed using θ^*, the KLIC minimized value of θ. When $\rho_1 \neq 0$, power depends on the sign of $\rho_1 b_1$, and power bounds are shown for positive and negative values of this product. For example, from panel A, when $\rho_1 = 0$ the power bound computed using $\theta^*$is 0.49 in the $I(1)$ null model with no time lags; power falls to 0.44 in the $I(1)$ model that allows for time lags, and continues to fall for more flexible trend specifications. In the unrestricted $g(s, r)$ model, the power bound is 0.36. The second set of entries in each panel show power bounds computed using the numerical approximations to the least favorable distributions for θ. As discussed in the appendix, these upper bounds are constructed so that they are no more than 2.5 percentage points above the actual least upper bounds (ignoring Monte Carlo integration error), and thus roughly correspond to the power of optimal tests for testing H_0 versus H_1. Results are shown for the $I(1)$ and local-to-unity models only. These models have one (ρ) and two (ρ, c) nuisance parameters, respectively, so that the computational algorithm described in the appendix can successfully be implemented. For the other specifications of the stochastic trend the number of nuisance parameters is large, and this makes it intractable to compute the least upper power bound.

The point optimal tests depend on the sign of $\rho_1 b_1$, and a researcher might be uncertain about this sign. Thus, it is interesting to examine tests designed to maximize weighted average power over alternatives with positive and negative values of $\rho_1 b_1$ (so that Γ in Theorem 1 has point mass at (ρ_1, b_1) and at $(\rho_1, -b_1)$). Of course, the optimal test will

depend on the relative weights given to these two alternatives. The third section of the tables shows power bounds for tests designed to maximize the minimum power over these two alternatives. For the cases considered in the Table, this always leads to a test with equal power against both these alternatives.

Glancing at the numbers in the tables, four results stand out. First, as noted by Stock and Watson (1996), Jansson and Moreira (2006), and others, the local-to-unity model shows asymmetric power functions when $\rho_1 \neq 0$. For example in the local-to-unity model with $|\rho_1| = 0.9$ and $|b_1| = 7$, the power bound is 0.48 for alternatives with $\rho_1 b_1 > 0$, but increases to 0.95 when $\rho_1 b_1 < 0$. This asymmetry is not present in the $I(1)$ model and also disappears for flexible versions of the null (the unrestricted model or stationary model with time lags). Second, in all models except the $I(1)$ null, the power bound associated with the two-sided test is essentially identical to the minimum of the point optimal tests corresponding to $\rho_1 b_1 < 0$ or $\rho_1 b_1 > 0$. That is, in terms of the minimum power associated with the two cases, the power bounds suggest that there is essentially no loss from using the two-sided test. The third result is that the power bounds associated with the KLIC minimized value of θ are only slightly larger than the (approximate) least upper power bounds computed for the $I(1)$ and local-to-unity models. Evidently then, in these cases the KLIC minimizers provide a good approximation to the least upper power bound.

Finally, the most important result in the table involves a comparison the power bounds to the power of the JW test. Recall that the JW test ignores X_T, uses only the data in Y_T, and tests whether Y_T is $I(0)$ against the local-level model alternative. The JW test is invariant and similar for H_0 versus H_1 so that its power cannot exceed the power bounds shown in the tables. Furthermore, its power does not depend on the value of ρ_1, so there is a single power value for the JW test for the alternatives considered in panel A, and single value for the alternatives in panel B. From (22), the JW test depends on $\bar{b}$, the value of b under the alternative, and results are presented for JW tests computed using the values of b_1 in panels A and B (denoted JW_{env}), and for the JW test computed using $\bar{b} = 8$. The power of the infeasible test JW_{env} is the power envelope of scale invariant tests of (21), while choosing $\bar{b} = 8$ produces a single feasible test (with critical values for the test statistic shown in Table 1). The power of the JW and JW_{env} tests are shown at the top of the panels. These numbers are very close, which indicates that there is only a small loss in power associated with not knowing b under the alternative. Comparing the power of the JW test to the

power bound for null model with an unrestricted stochastic trend (the column labeled UNR) indicates that the JW test is, for all practical purposes, an optimal test. That is, the JW test is (essentially) efficient among the class of invariant tests that control size uniformly over $g(s,r)$ in the unrestricted model. Moreover, when ρ_1 is small, their are only small potential power gains associated with the restriction that $g(s,r)$ is stationarity or, when time lags are allowed, that $g(s,r)$ decays exponentially as in the local-to-unity models. Indeed, the difference between the power bounds shown in the table and the power of the JW test is large only when the null is highly restricted (such as the $I(1)$ model), or $|\rho_1|$ is large and $\rho_1 b_1 < 0$ (such as the stationary null model with no time lags and $|\rho_1| = 0.9$), and even then, the test JW is often close to admissible in the sense that the two-sided tests only have marginally higher power.

8 Conclusions

This paper studies inference about the cointegrating vector in a bivariate framework that uses a flexible model for the common stochastic trend (incorporating the $I(1)$, local-to-unity, and fraction model as special cases) and that allows considerable flexibility in the timing of interactions of the trend and error correction term. Inference is carried out using the low-frequency transformations of the data suggested by Müller and Watson (2006) in a study of univariate models. This paper develops a general method for computing upper bounds on the power of asymptotically valid tests. The method is used to derive bounds on the power of tests that control size over flexible stochastic trend specifications, and which maximize power against alternatives with the usual $I(1)$ trend. We find that a low-frequency version of Wright's (2000) test almost achieves this upper bound on power.

The implication for applied work is that, at least in a bivariate framework, approximately efficient robust inference about the cointegrating vector may be carried out using this test. The test is robust in two ways. First, it is robust to arbitrary autocorrelation properties in the error correction term above a pre-specified low-frequency band. Second, it is robust to the precise nature of persistence, as its rejection probability under the null hypothesis does not depend on the nature of the stochastic trend. Thus, in absence of precise *a prior* information about the nature of the stochastic trend, the test is essentially efficient, as no test exists with substantially higher power.

The test is straightforward to implement: Section 5 above provides a simple version of the associated test statistic and tabulates asymptotically justified critical values. As in Wright (2000), confidence sets for the cointegrating vector are then easily obtained be inverting the test.

A Appendix

A.1 Proof of Lemma 1

We first establish a preliminary result.

Lemma 4 *For any $0 \leq r < 1$, the functions $\Psi_l : [0,r] \mapsto \mathbb{R}$ with $\Psi_l(s) = \sqrt{2}\cos(\pi l s)$, $l = 1, \cdots, 2q$ are linearly independent.*

Proof. Choose any real constants c_j, $j = 1, \cdots, 2q$, so that $\sum_{j=1}^{2q} c_j \Psi_j(s) = 0$ for all $s \in [0,r]$. Then also $\sum_{j=1}^{2q} c_j \Psi_j^{(k)}(0) = 0$ for all $k > 0$, where $\Psi_j^{(k)}(0)$ is the kth (right) derivative of $\Psi_j(s)$ at $s = 0$. A direct calculation shows $\Psi_j^{(k)}(0) = -(-1)^{k/2}\sqrt{2}(\pi j)^k$ for even k. It is not hard to see that the $2q \times 2q$ matrix with j,lth element $-(-1)^{k/2}(\pi j)^k$ is nonsingular, so that $\sum_{j=1}^{2q} c_j \Psi_j^{(k)}(0) = 0$ can only hold for $c_j = 0$, $j = 1, \cdots, 2q$. ■

To begin the proof to Lemma 1 write

$$\Sigma^* = \begin{pmatrix} I_q & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

and note that since Σ^* is positive definite, so is $I_q - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Therefore, there exists $\varepsilon > 0$ such that with $\tilde{Z}_l = \int_0^{1-\varepsilon} \Psi_l(r) dW_v(r)$, $E[\tilde{Z}\tilde{Z}'] - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ is positive definite.

Let $\tilde{\Psi}_{q+1}(s)$ be the scaled residuals of continuous time projection of $\mathbf{1}[s \leq 1-\varepsilon]\Psi_{q+1}(s)$ on $\{\mathbf{1}[s \leq 1-\varepsilon]\Psi_l(s)\}_{l=1}^q$ on the unit interval, and let $\tilde{\Psi}_{q+j}(s)$, $j = 2, \cdots, q$ be the scaled residuals of continuous time projection of $\mathbf{1}[s \leq 1-\varepsilon]\Psi_{q+j}(s)$ on $\{\mathbf{1}[s \leq 1-\varepsilon]\Psi_l(s)\}_{l=1}^q$ and $\{\mathbf{1}[s \leq 1-\varepsilon]\tilde{\Psi}_{q+l}(s)\}_{l=1}^{j-1}$. By Lemma 4, $\tilde{\Psi}_j(s)$, $j = q+1, \cdots, 2q$, are not identically zero, and we can choose their scale to make them orthonormal. Let $\zeta_l = \int_0^1 \tilde{\Psi}_{q+l}(r) dW_v(r)$, $l = 1, \cdots, q$, so that by construction, $(\tilde{Z}', \zeta')' \sim \mathcal{N}(0, \mathrm{diag}(E[\tilde{Z}\tilde{Z}'], I_q))$. Because the $2q \times 2q$ matrix

$$\tilde{\Sigma} = \begin{pmatrix} E[\tilde{Z}\tilde{Z}'] & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

is positive definite there exist $q^2+q(q+1)/2$ numbers $p_{l,j}$, $l=1,\cdots,q$, $j=1,\cdots,q+l$, for which

$$\tilde{V}_l = \sum_{j=1}^{q} p_{l,j}\tilde{Z}_j + \sum_{j=1}^{l} p_{l,q+j}\zeta_j \text{ for } l=1,\cdots,q$$

is distributed $(\tilde{Z}',\tilde{V}')' \sim \mathcal{N}(0,\tilde{\Sigma})$. Note that by construction, $(Z',\tilde{V}')' \sim \mathcal{N}(0,\Sigma^*)$, since $(Z-\tilde{Z})$ and $(\tilde{Z}',\tilde{V}')'$ are independent. By the linearity of Ito-integrals, we can rewrite

$$\tilde{V}_l = \int_0^1 \tilde{f}_l(r)dW_v(r)$$

where $\tilde{f}_l(r)=0$ for $r>1-\varepsilon$ and

$$\tilde{f}_l(r) = \sum_{j=1}^{q} p_{l,j}\Psi_j(r) + \sum_{j=1}^{l} p_{l,q+j}\tilde{\Psi}_{q+l}(r)$$

for $r\in[0,1-\varepsilon]$. Since $\Psi_l(s)=(-1)^l\Psi_l(1-s)$ for all $l\geq 1$, Lemma 4 implies that the $q\times q$ matrix $H(r)=\int_r^1 \Psi(s)\Psi(s)'ds$ is nonsingular for all $r\leq 1-\varepsilon$. Now set $g(s,r)=0$ for $r\in(-\infty,0)\cup(1-\varepsilon,1]$ and $g(s,r)=\Psi(s)'H(r)^{-1}\tilde{f}(r)$ for $r\in[0,1-\varepsilon]$. Then for $0\leq s<t\leq 1$, we have with $\psi(s)=d\Psi(s)/ds$

$$\begin{aligned}
& E[|\int_0^t g(t,r)dW_v(r) - \int_0^s g(s,r)dW_v(r)|^4] \\
= \;& 3[\int_0^s (g(t,r)-g(s,r))^2dr + \int_s^t g(t,r)^2dr]^2 \\
\leq \;& 3[(\sup_{0\leq r\leq 1}||H(r)^{-1}||^2||\tilde{f}(r)||^2)(||\Psi(s)-\Psi(t)||^2 + (t-s)\sup_{0\leq r\leq 1}||\Psi(r)||^2)]^2 \\
\leq \;& 3(\sup_{0\leq r\leq 1}||H(r)^{-1}||^4||\tilde{f}(r)||^4)(\sup_{0\leq r\leq 1}||\psi(r)||^2 + \sup_{0\leq r\leq 1}||\Psi(r)||^2)^2(t-s)^2
\end{aligned}$$

where the last inequality follows from $\Psi(t)-\Psi(s)=\int_0^1(t-s)\psi(s+\lambda(t-s))d\lambda$, so that by Kolmogorov's continuity theorem, there exists a continuous (and thus cadlag) version of the stochastic process $\int_0^s \Psi(s)'H(r)^{-1}\tilde{f}(r)dW_v(r)$. Furthermore, $V=\int_0^1\int_r^1 \Psi(s)g(s,r)ds\,dW_v(r) = \int_0^{1-\varepsilon}\int_r^1 \Psi(s)\Psi(s)'H(r)^{-1}\tilde{f}(r)ds\;dW_v(r)=\tilde{V}$ almost surely.

A.2 Proof of Lemma 2

Note that Q is a maximal invariant of the group of transformations

$$\begin{bmatrix} Y \\ X \end{bmatrix} \to \begin{bmatrix} a_{yy}Y \\ a_{xy}Y + a_{xx}X \end{bmatrix}$$

for non-zero a_{yy} and a_{xx}, and arbitrary a_{xy}. This group of transformation induces a corresponding group on the parameter space

$$(b, \gamma_1, \gamma_2, \gamma_3, \theta) \rightarrow (b, a_{yy}\gamma_1, a_{xy}\gamma_1 + a_{xx}\gamma_2, a_{xx}\gamma_3, \theta). \tag{25}$$

By Theorem 6.3.2 of Lehmann and Romano (2005), the distribution of Q only depends on the parameters through a maximal invariant of (25), so the distribution of Q does not depend on γ_1, γ_2 and γ_3. We might thus set $\gamma_1 = \gamma_3 = 1$ and $\gamma_2 = 0$. The $2q \times 1$ vector

$$(Q', Y_q, \frac{X_q}{Y_q}, X_{q-1} - \frac{X_q}{Y_q}Y_{q-1})'$$

is a one-to-one mapping from (Y, X), so that its density in terms of f_θ is easily computed. The first result now follows by computing the marginal density of Q.

Furthermore, by a change of variable

$$\begin{aligned} & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\gamma_1\gamma_3|^{-q+1} f_\theta(\gamma_1^{-1}Y, \gamma_3^{-1}(X - \gamma_2 Y))d\gamma_1 d\gamma_2 d\gamma_3 \\ = & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |s_1|^{q-1}|s_3|^{q-2} f_\theta(s_1 Y, s_3 X + s_2 Y) ds_2 ds_1 ds_3. \end{aligned}$$

Since f is the density of a multivariate normal with covariance matrix $\Sigma(b)$, we find

$$\begin{aligned} & -2\ln f_\theta(s_1 Y, s_3 X + s_2 Y) + C \\ = \; & s_1^2(Y'\Sigma(b)^-_{11}Y) + 2s_1 s_3(Y'\Sigma(b)^-_{12}X) + s_3^2(X'\Sigma(b)^-_{22}X) \\ & +2s_2[s_1(Y'\Sigma(b)^-_{12}Y) + s_3(X'\Sigma(b)^-_{22}Y)] + s_2^2(Y'\Sigma(b)^-_{22}Y) \end{aligned}$$

where C is a generic constant that does not depend on Y and X. By 'completing the square' for s_2, we thus find

$$\begin{aligned} & \int_{-\infty}^{\infty} |s_1|^{q-1}|s_3|^{q-2} f_\theta(s_1 Y, s_3 X + s_2 Y) ds_2 \\ = \; & C(Y'\Sigma^-_{i,22}Y)^{-1/2} \exp[-\tfrac{1}{2}(Y'\Sigma^-_{i,22}Y)^{-1}(A_X s_3^2 + A_Y s_1^2 + 2A_{XY}s_1 s_3)]. \end{aligned}$$

But

$$\begin{aligned} & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |s_1|^{q-1}|s_3|^{q-2} \exp[-\tfrac{1}{2}(Y'\Sigma^-_{i,22}Y)^{-1}(A_X s_3^2 + A_Y s_1^2 + 2A_{XY}s_1 s_3)] ds_1 ds_3 \\ = \; & CE[|S_1|^{q-1}|S_2|^{q-2}] \end{aligned}$$

where S_1 and S_2 are bivariate normal with variances $A_Y(Y'\Sigma_{i,22}^{-}Y)/(A_YA_X - A_{XY}^2)$ and $A_X(Y'\Sigma_{i,22}^{-}Y)/(A_YA_X - A_{XY}^2)$, respectively, and correlation coefficient $A_{XY}/\sqrt{A_YA_X}$. The result now follows from Nabeya's (1951) formula for the absolute moments of a bivariate normal and some rearranging.

A.3 Proof of Theorem 1

(i) Since φ is a bounded and almost everywhere continuous function, and the limiting distribution of Q is absolutely continuous, $Q_T \Rightarrow Q$ implies $E\varphi(Q_T) \to E\varphi(Q)$.

(ii) Let $F_{1,T}$ be the distribution of Q_T under the alternative model M_1, where b and θ are drawn according to Γ, so that for any bounded and continuous function $\vartheta : \mathbb{R}^{2q-3} \mapsto \mathbb{R}$, $\int \vartheta dF_{1,T} \to \int \vartheta dF_{1,\infty} = \int \vartheta h dQ$. Define for some $\bar{K} > 0$ the indicator function $\mathcal{B}$ of the set $B = \{x \in \mathbb{R}^{2q-3} : \mathrm{LR}(x) \leq \bar{K}\}$, where $\mathrm{LR} = \int f_Q(M_0, 0, \theta)d\Lambda(\theta)/h$. Define the distribution $F_{0,\infty}$ as $\int \vartheta dF_{0,\infty} = \int \vartheta \int f_Q(M_0, 0, \theta)d\Lambda(\theta)dQ$, and let $F_{0,T}$ be the probability distribution defined via

$$\int \vartheta dF_{0,T} = \int \mathcal{B} dF_{0,\infty} \frac{\int \vartheta \mathcal{B}\,\mathrm{LR}\, dF_{1,T}}{\int \mathcal{B}\,\mathrm{LR}\, dF_{1,T}} + \int \vartheta(1-\mathcal{B})dF_{0,\infty}.$$

Note that $\int \vartheta dF_{0,T} \to \int \mathcal{B}\vartheta\,\mathrm{LR}\, dF_{1,\infty} + \int \vartheta(1-\mathcal{B})dF_{0,\infty} = \int \vartheta dF_{0,\infty}$, where the convergence follows because $\mathcal{B}$ and $\mathcal{B}\vartheta\,\mathrm{LR}$ are bounded and almost everywhere continuous functions, and the limiting distribution $F_{1,\infty}$ is absolutely continuous. The distribution $F_{0,T}$ hence converges weakly to the distribution of a limiting random vector Q under the null model M_0, where $\theta \in \Theta_0$ is drawn according to Λ. Thus, (19) implies that $\limsup_{T\to\infty} \int \varphi_T dF_{0,T} \leq \alpha$. By the Neyman-Pearson Lemma, for any T, the most powerful test to discriminate between the distributions $F_{0,T}$ and $F_{1,T}$, conditional on $Q_T \in B$, is given by a test that rejects for small values of LR. But the test φ_Λ also rejects for small values of LR, and satisfies $\limsup_{T\to\infty} \int \varphi_\Lambda dF_{0,T} \leq \alpha$ by applying part (i) of the Theorem. Thus, even with perfect discrimination between $F_{0,T}$ and $F_{1,T}$ for $Q_T \in B$, $\limsup_{T\to\infty} \mathrm{WAP}_T(\varphi_T) \leq \limsup_{T\to\infty}(\mathrm{WAP}_T(\varphi_\Lambda) + \int(1-\mathcal{B})dF_{1,T}) = \int \vartheta h dQ + \int(1-\mathcal{B})dF_{1,\infty}$, at least as long $\bar{K} > 1/k_\Lambda$. The result now follows after observing that $\int(1-\mathcal{B})dF_{1,\infty}$ can be made arbitrarily small by choosing $\bar{K}$ large.

A.4 Algorithm for approximating the least favorable distribution, least upper power bound, and optimal test

By Theorem 1, it suffices to consider exact small sample hypothesis testing problem concerning Q, i.e. (20). For some integer $N > 0$, let $\bar{\theta}_N = (\theta_1, \cdots, \theta_N)$ and $\bar{p}_N = (p_1, \cdots, p_N)$, where $\theta_i \in \Theta_0$ and $p_i \geq 0$ for all i, and $\sum_{i=1}^N p_i = 1$. For $x \in \mathbb{R}^{2q-3}$, define the test $\varphi(\bar{\theta}, \bar{p}, k)$ as

$$\varphi(\bar{\theta}, \bar{p}, k)(x) = \mathbf{1}[\frac{h(x)}{\sum_{i=1}^N p_i f_Q(M_0, 0, \theta_i)(x)} > k]$$

where we write $h(x)$ and $f_Q(M_0, 0, \theta_i)(x)$ for the densities h and $f_Q(M_0, 0, \theta_i)$ evaluated at x. Introduce the notation $\Pi_0(\bar{\theta}, \bar{p}, k; \theta) = \int \varphi(\bar{\theta}, \bar{p}, k) f_Q(M_0, 0, \theta) dQ$ for the rejection probability of the test $\varphi(\bar{\theta}, \bar{p}, k)$ under the null hypothesis with $\theta \in \Theta_0$, and $\Pi_1(\bar{\theta}, \bar{p}, k) = \int \varphi(\bar{\theta}, \bar{p}, k) h dQ$ for its rejection probability under the alternative. The algorithm requires repeated evaluations of $\Pi_0(\bar{\theta}, \bar{p}, k; \theta)$ and $\Pi_1(\bar{\theta}, \bar{p}, k)$. By Lemma 2, it is possible to express $f_Q(M, b, \theta)(Q)$ in terms of the $2q \times 1$ mean zero multivariate normal $(Y', X')'$ with covariance matrix (12), and by Lemma 2, the density of Q does not depend on the value of γ. We thus choose $\gamma_1 = \gamma_3 = 1$ and $\gamma_2 = 0$ without loss of generality. Let ξ_j, $j = 1, \cdots, n$ be independent pseudo random vectors with distribution $\xi_j \sim N(0, I_{2q})$. We numerically approximate $\Pi_0(\bar{\theta}, \bar{p}, k; \theta)$ by

$$\hat{\Pi}_0(\bar{\theta}, \bar{p}, k; \theta) = \frac{1}{n} \sum_{j=1}^n \Upsilon_L \left(\frac{h(\hat{Q}_j(\theta))}{\sum_{i=1}^N p_i f_Q(M_0, 0, \theta_i)(\hat{Q}_j(\theta))}, k \right)$$

where for some $L > 0$, $\Upsilon_L : \mathbb{R}^2 \mapsto \mathbb{R}$ is defined as $\Upsilon(x, k) = x^L/(k^L + x^L)$ and $\hat{Q}_j(\theta)$ is defined as Q in (15) with Y and X replaced by the first q and last q elements of the $2q \times 1$ vector $C(\theta)\xi_j$, with $C(\theta)$ the Cholesky decomposition of $\Sigma_{(Y,X)}$ of model M_0 with parameter θ and $b = 0$.Define $\hat{\Pi}_1(\bar{\theta}, \bar{p}, k)$ analogously. The pseudo random vectors ξ_j, $j = 1, \cdots, n$, are only drawn once in the evaluation of $\hat{\Pi}_0(\bar{\theta}, \bar{p}, k; \theta)$ and $\hat{\Pi}_1(\bar{\theta}, \bar{p}, k)$ at different arguments. Note that as $L \to \infty$, $\Upsilon(x, k) \to \mathbf{1}[x > k]$, so that for L large, $\hat{\Pi}_0(\bar{\theta}, \bar{p}, k; \theta)$ approximates the standard Monte Carlo approximation for the rejection probability of $\varphi(\bar{\theta}, \bar{p}, k)$. The advantage of choosing $L < \infty$ is that $\hat{\Pi}_0(\bar{\theta}, \bar{p}, k; \theta)$ and $\hat{\Pi}_1(\bar{\theta}, \bar{p}, k)$ become smooth and differentiable functions of their arguments, which greatly simplifies numerical maximizations. The computations in Table 2 were performed with n =25,000 and $L = 9$.

The output of the algorithm is an approximate discrete least favorable distribution Λ^*, described by $(\bar{\theta}^*, \bar{p}^*)$, where Λ^* puts mass p_i^* on the points θ_i^* for $i = 1, \cdots, N^*$. The

distribution Λ^* is approximately least favorable in the sense that the level α likelihood ratio test $\varphi(\bar{\theta}^*, \bar{p}^*, k_\Lambda^*)$ of $H_{\Lambda^*} : Q$ has density $\int f_Q(M_0, 0, \theta) d\Lambda^*(\theta)$ against H_1 has (approximately) $\varepsilon_\Pi > 0$ more power than the critical value adjusted test $\varphi(\bar{\theta}^*, \bar{p}^*, k_0^*)$ of H_0 against H_1 of size smaller or equal to α. By construction, the (approximate) power bound $\hat{\Pi}_1(\bar{\theta}^*, \bar{p}^*, k^*)$ is no more than ε_Π above the least upper power bound, and the valid level α test $\varphi(\bar{\theta}^*, \bar{p}^*, k_0^*)$ of H_0 against H_1 has power that is at most ε_Π below the power bound. We set $\varepsilon_\Pi = 0.025$ in the computations for $\hat{\Pi}_1(\bar{\theta}^*, \bar{p}^*, k^*)$ of Table 2.

The algorithm consists of two main subroutines, say SR I and SR II. SR I takes $(\bar{\theta}, \bar{p})$ as given and either identifies $(\bar{\theta}, \bar{p})$ as yielding Λ^*, or returns $\hat{\theta} \in \Theta_0$. SR I performs the following computations:

1. Solve for the real number k_Λ that satisfies $\sum_{i=1}^N p_i \hat{\Pi}_0(\bar{\theta}, \bar{p}, k_\Lambda; \theta_i) = \alpha$, so that the test $\varphi(\bar{\theta}, \bar{p}, k_\Lambda)$ is (approximately) the level α likelihood ratio test of $H_\Lambda : Q$ has density $\int f_Q(M_0, 0, \theta) d\Lambda(\theta)$, where Λ is a discrete distribution that puts mass p_i on the points θ_i.

2. Compute $\hat{\Pi}_1(\bar{\theta}, \bar{p}, k_\Lambda)$, and solve for $k_\Lambda^c > k_\Lambda$ such that $\hat{\Pi}_1(\bar{\theta}, \bar{p}, k_\Lambda) - \hat{\Pi}_1(\bar{\theta}, \bar{p}, k_\Lambda^c) = \varepsilon_\Pi$.

3. Numerically maximize $\hat{\Pi}_0(\bar{\theta}, \bar{p}, k_\Lambda^c; \theta)$ over θ. Denote the argmax by $\hat{\theta}$. If $\hat{\Pi}_0(\bar{\theta}, \bar{p}, k_\Lambda^c; \hat{\theta}) \leq \alpha$, then the test $\varphi(\bar{\theta}, \bar{p}, k_\Lambda^c)$ is a valid level α test of H_0 against H_1, and its power $\hat{\Pi}_1(\bar{\theta}, \bar{p}, k_\Lambda^c)$ differs by only ε_Π from the power upper bound $\hat{\Pi}_1(\bar{\theta}, \bar{p}, k_\Lambda)$, so that Λ^* is identified. If $\hat{\Pi}_0(\bar{\theta}, \bar{p}, k_\Lambda^c; \hat{\theta}) \geq \alpha$, return $\hat{\theta}$.

SR II takes $\bar{\theta} = (\theta_1, \cdots, \theta_N)$ as given, and returns a new set $(\bar{\theta}^n, \bar{p}^n)$ with $\bar{\theta}^n = (\theta_1^n, \cdots, \theta_{N^n}^n)$ and $\bar{p} = (p_1^n, \cdots, p_{N^n}^n)$, with N^n possibly different from N, and proceeds by the following steps:

1. Numerically identify a preliminary $\bar{p}^p = (p_1^p, \cdots, p_N^p)$ with $p_i^p \geq 0$ for all i and $\sum_{i=1}^N p_i^p = 1$ and real number k^p such that

$$\hat{\Pi}_0(\bar{\theta}, \bar{p}^p, k^p; \theta_i) \leq \alpha \quad \text{and} \quad p_i(\hat{\Pi}_0(\bar{\theta}, \bar{p}^p, k^p; \theta_i) - \alpha) = 0 \ \text{ for } i = 1, \cdots, N. \tag{26}$$

   The $N \times 1$ vector $\bar{p}^p$ is (a numerical approximation of) the least favorable distribution in the level α hypothesis testing problem of $H_{\bar{\theta}}$ against H_1, where under $H_{\bar{\theta}}$, the density of Q is known to be one of $f_Q(M_0, 0, \theta_i)$, $i = 1, \cdots, N$, and p_i^p describes the least

favorable probability that Q has density $f_Q(M_0, 0, \theta_i)$. By Theorem 3.8.1 of Lehmann and Romano (2005), the resulting test $\varphi(\bar{\theta}, \bar{p}^p, k^p)$ has the two properties that (i) $\Pi_0(\bar{\theta}, \bar{p}^p, k^p; \theta_i) \leq \alpha$ for $i = 1, \cdots, N$ and (ii) $\Pi_0(\bar{\theta}, \bar{p}^p, k^p; \theta_i) < \alpha$ only if $p_i^p = 0$. It is numerically convenient to use these properties of the least favorable distribution directly in the numerical approximation, leading to (26).

2. Collect the nonzero elements in $\bar{p}^p$ in a new N^n-dimensional vector $\bar{p}^n$, and collect the corresponding elements of $\bar{\theta}^p$ in $\bar{\theta}^n$.

Overall the algorithm iterates between the two blocks as follows:

1. Set $j = 0$, and initialize by letting $N_{(0)} = \bar{p}_{(0)} = 1$ and $\bar{\theta}_{(0)} = \theta^*$, where θ^* is the numerical minimizer of the Kulback-Leibler divergence problem (24).

2. Call SR I with $(\bar{\theta}, \bar{p}) = (\bar{\theta}_{(j)}, \bar{p}_{(j)})$. If SR I identifies $(\bar{\theta}_{(j)}, \bar{p}_{(j)})$ as describing Λ^*, the algorithm stops.

3. Otherwise, call SR II with $\bar{\theta} = (\bar{\theta}_{(j)}, \hat{\theta})$, where $\hat{\theta}$ is the value returned by SR I in the last call.

4. Increase j by one, set $(\bar{\theta}_{(j)}, \bar{p}_{(j)}) = (\bar{\theta}^n, \bar{p}^n)$, where $(\bar{\theta}^n, \bar{p}^n)$ are the values returned by the last call of SR II, and go to Step 2.

In words, the algorithm may be described as follows: SR I determines whether the discrete distribution $\Lambda_{(j)}$ described by the current nodes $(\bar{\theta}_{(j)}, \bar{p}_{(j)})$ already represents Λ^* by searching for a value of θ that induces a too large rejection probability (in the sense that a small adjustment of the critical of the level α test $\varphi_{\Lambda_{(j)}}$ of $H_{\Lambda_{(j)}}$ against H_1 is not enough to yield a test of size α of H_0 against H_1). If a large rejection probability can be found, then the value $\hat{\theta}$ which induces this overrejection describes a density $f_Q(M_0, 0, \hat{\theta})$ which is "too similar" to the density h under H_1. The value $\hat{\theta}$ should therefore be included in the list of nodes. SR II appropriately reweighs the probability masses in this new list of nodes, and eliminates superfluous nodes. Then SR I is again called to check whether the new list of nodes yields Λ^*, etc.

The first step of SR II looks computationally daunting, but this is not so: given that $\bar{\theta}$ is fixed, one can compute the $(N+1)Nn$ numbers $h(\hat{Q}_j(\theta_l))$ and $f_Q(M_0, 0, \theta_i)(\hat{Q}_j(\theta_l))$ for $i, l =$

$1, \cdots, N$ and $j = 1, \cdots, n$ once and then use those directly to solve (26). Furthermore, since $\hat{\Pi}_0(\bar{\theta}, \bar{p}, k; \theta)$ is differentiable with respect to its arguments, (26) can be well approximated by the minimization of a suitably defined, differentiable objective function, so that gradient methods can be employed. The computationally most demanding aspect of the algorithm is Step 3 of SR I. One can again use gradient methods in the maximization, but there is of course no guarantee that the problem is globally concave in θ. For the computations in Table 2, we used 100 different starting values for θ in Step 3 of SR I, and at most 200 gradient steps. The starting values for c in the local-to-unity model where chosen on the interval $(-3, 100)$, and values for (ρ, c) that lead to a condition number of $\Sigma_{(Y,X)}$ above 10^5 were excluded from the numerical maximization in Step 3 of SR I.

References

Bierens, H. J. (1997): "Nonparametric Cointegration Analysis," *Journal of Econometrics*, 77, 379–404.

Bobkoski, M. J. (1983): "Hypothesis Testing in Nonstationary Time Series," *unpublished Ph.D. thesis, Department of Statistics, University of Wisconsin*.

Campbell, J. Y., and M. Yogo (2006): "Efficient Tests of Stock Return Predictability," *Journal of Financial Economics*, 81, 27–60.

Canjels, E. (1997): "Essays on Time Series Econometrics and Sharecropping in Agriculture," *Ph.D. Thesis, Northwestern University*.

Cavanagh, C. L. (1985): "Roots Local To Unity," *Working Paper, Harvard University*.

Cavanagh, C. L., G. Elliott, and J. H. Stock (1995): "Inference in Models with Nearly Integrated Regressors," *Econometric Theory*, 11, 1131–1147.

Chan, N. H., and C. Z. Wei (1987): "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *The Annals of Statistics*, 15, 1050–1063.

Elliott, G. (1998): "The Robustness of Cointegration Methods When Regressors Almost Have Unit Roots," *Econometrica*, 66, 149–158.

Elliott, G., and U. K. Müller (2006): "Efficient Tests for General Persistent Time Variation in Regression Coefficients," *Review of Economic Studies*, 73, 907–940.

Elliott, G., and J. H. Stock (1994): "Inference in Time Series Regression When the Order of Integration of a Regressor is Unknown," *Econometric Theory*, 10, 672–700.

Granger, C. W. J. (1993): "What are We Learning Abou the Long-Run?," *The Economic Journal*, 103, 307–317.

Jansson, M., and M. J. Moreira (2006): "Optimal Inference in Regression Models with Nearly Integrated Regressors," *Econometrica*, 74, 681–714.

JEGANATHAN, P. (1997): "On Asymptotic Inference in Linear Cointegrated Time Series Systems," *Econometric Theory*, 13, 692–745.

——— (1999): "On Asymptotic Inference in Cointegrated Time Series with Fractionally Integrated Errors," *Econometric Theory*, 15, 583–621.

JOHANSEN, S. (1988): "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.

KIM, C. S., AND P. C. B. PHILLIPS (2000): "Fully Modified Estimation of Fractional Cointegration Models," *Working Paper, Yale University.*

KWIATKOWSKI, D., P. C. B. PHILLIPS, P. SCHMIDT, AND Y. SHIN (1992): "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root," *Journal of Econometrics*, 54, 159–178.

LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*. Springer, New York.

MARINUCCI, D., AND P. M. ROBINSON (1999): "Alternative Forms of Fractional Brownian Motion," *Journal of Statistical Planning and Inference*, 80, 111–122.

MÜLLER, U. K. (2006): "A Theory of Robust Long-Run Variance Estimation," *Working paper, Princeton University.*

MÜLLER, U. K., AND M. W. WATSON (2006): "Testing Models of Low-Frequency Variability," *NBER Working Paper W12671.*

NABEYA, S. (1951): "Absolute Moments in 2-Dimensional Normal Distribution," *Annals of the Institute of Statistical Mathematics*, 3, 2–6.

NYBLOM, J. (1989): "Testing for the Constancy of Parameters Over Time," *Journal of the American Statistical Association*, 84, 223–230.

PARK, J. Y. (1992): "Canonical Cointegrating Regressions," *Econometrica*, 60, 119–143.

PHILLIPS, P. C. B. (1987): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547.

——— (1998): "New Tools for Understanding Spurious Regression," *Econometrica*, 66, 1299–1325.

Phillips, P. C. B., and B. E. Hansen (1990): "Statistical Inference in Instrumental Variables Regression with I(1) Processes," *Review of Economic Studies*, 57, 99–125.

Robinson, P. M., and J. Hualde (2003): "Coinegration in Fractional Systems with Unknown Integration Orders," *Econometrica*, 71, 1727–1766.

Robinson, P. M., and D. Marinucci (2003): "Semiparametric Frequency Domain Analysis of Fractional Cointegration," in *Time Series with Long Memory*, ed. by P. M. Robinson, pp. 334–373. Oxford University Press, Oxford.

Saikkonen, P. (1991): "Estimation and Testing of Cointegrating Regressions," *Econometric Theory*, 7, 1–27.

Sriananthakumar, S., and M. L. King (2006): "A New Approximate Point Optimal Test of a Composite Null Hypothesis," *Journal of Econometrics*, 130, 101–122.

Stock, J. H., and M. W. Watson (1993): "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, 61, 783–820.

——— (1996): "Confidence Sets in Regressions with Highly Serially Correlated Regressors," *Working Paper, Harvard University*.

Velasco, C. (2003): "Gaussian Semiparametric Estimation of Fractional Cointegration," *Journal of Time Series Analysis*, 24, 345–378.

Wright, J. H. (2000): "Confidence Sets for Cointegrating Coefficients Based on Stationarity Tests," *Journal of Business and Economic Statistics*, 18, 211–222.

Table 1: Critical Values for the JW Statistic

q	**Size**		
	10%	**5%**	**1%**
6	2.48	2.89	4.01
7	2.21	2.52	3.37
8	2.02	2.27	2.97
9	1.87	2.09	2.65
10	1.76	1.95	2.44
11	1.68	1.84	2.26
12	1.61	1.75	2.13
13	1.55	1.68	2.02
14	1.51	1.62	1.92
15	1.47	1.57	1.85
16	1.43	1.53	1.78
17	1.40	1.49	1.72
18	1.38	1.46	1.67

Note: The table shows asymptotic critical for the JW statistic computed using $\bar{b} = 8$.

Table 2: Power Bounds for Cointegrating Coefficient Tests

H_1: $b = b_1$, $\rho = \rho_1$, $v_t \sim I(1)$

A. $|b_1| = 7$ (Power of JW Test = 0.36, Power of JW_{env} Test = 0.36)

	Restrictions on Nuisance Parameters in Null Model						
	I(1)	I(1) – TL	LTU	LTU - TL	STA	STA – TL	UNR
(i) KLIC Minimized Point Optimal Tests							
$\|\rho_1\| = 0.0$	0.49	0.44	0.49	0.39	0.41	0.37	0.36
$\|\rho_1\| = 0.5$	0.64 0.64	0.59 0.59	0.51 0.65	0.42 0.46	0.39 0.52	0.39 0.38	0.36 0.36
$\|\rho_1\| = 0.9$	0.95 0.95	0.93 0.93	0.48 0.95	0.47 0.92	0.43 0.89	0.43 0.42	0.36 0.36
(ii) Approximate Least Favorable Distribution Point Optimal Tests							
$\|\rho_1\| = 0.0$	0.49		0.49				
$\|\rho_1\| = 0.5$	0.64 0.64		0.51 0.64				
$\|\rho_1\| = 0.9$	0.94 0.94		0.48 0.94				
(iii) 2-Sided Tests							
$\|\rho_1\| = 0.0$	0.49	0.44	0.49	0.39	0.41	0.37	0.36
$\|\rho_1\| = 0.5$	0.55	0.54	0.51	0.42	0.39	0.38	0.36
$\|\rho_1\| = 0.9$	0.90	0.88	0.48	0.47	0.43	0.42	0.36

B. $|b_1| = 14$ (Power of JW Test = 0.62, Power of JW_{env} Test = 0.63)

	Restrictions on Nuisance Parameters in Null Model						
	I(1)	I(1) – TL	LTU	LTU - TL	STA	STA – TL	UNR
(i) KLIC Minimized Point Optimal Tests							
$\|\rho_1\| = 0.0$	0.81	0.79	0.79	0.66	0.69	0.65	0.64
$\|\rho_1\| = 0.5$	0.90 0.90	0.88 0.88	0.76 0.91	0.70 0.68	0.67 0.80	0.67 0.67	0.64 0.64
$\|\rho_1\| = 0.9$	1.00 1.00	1.00 1.00	0.74 1.00	0.73 0.88	0.72 0.98	0.71 0.71	0.65 0.66
(ii) Approximate Least Favorable Distribution Point Optimal Tests							
$\|\rho_1\| = 0.0$	0.81		0.77				
$\|\rho_1\| = 0.5$	0.90 0.90		0.74 0.88				
$\|\rho_1\| = 0.9$	1.00 1.00		0.71 1.00				
(iii) 2-Sided Tests							
$\|\rho_1\| = 0.0$	0.81	0.79	0.79	0.66	0.69	0.64	0.64
$\|\rho_1\| = 0.5$	0.86	0.85	0.76	0.67	0.67	0.66	0.64
$\|\rho_1\| = 0.9$	1.00	0.99	0.74	0.73	0.72	0.71	0.65

Notes: Entries are power bounds for 5% tests. See the text for a description of the tests. The restrictions on the null model are: the $I(1)$ model without time lags (labeled $I(1)$), the $I(1)$ model with time lags ($I(1)$-TL), the local-to-unity model with and without time lags (LTU and LTU-TL, respectively), the stationary model with and without time lags (STA and STA-TL, respectively), and the unrestricted model (UNR). Cells containing two entries show power for the alternative $\rho_1 b_1 > 0$ and $\rho_1 b_1 < 0$, respectively.