

# Estimation and Hypothesis Testing of Cointegration

by

Adam Nowak

A Dissertation Presented in Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy

Approved October 2011 by the  
Graduate Supervisory Committee:

Seung Ahn, Chair  
Jarl Kallberg  
Crocker Liu

ARIZONA STATE UNIVERSITY

December 2012

## ABSTRACT

Estimating cointegrating relationships requires specific techniques. Canonical correlations are used to determine the rank and space of the cointegrating matrix. The vectors used to transform the data into canonical variables have an eigenvector representation, and the associated canonical correlations have an eigenvalue representation. The number of cointegrating relations is chosen based upon a theoretical difference in the convergence rates of the eigenvalues. The number of cointegrating relations is consistently estimated using a threshold function which places a lower bound on the eigenvalues associated with cointegrating relations and an upper bound on the eigenvalues on the eigenvalues not associated with cointegrating relations. The proposed estimator performs better with a large number of cross-sectional observations and moderate time series length.

## TABLE OF CONTENTS

	Page
LIST OF TABLES.....	iv
LIST OF FIGURES.....	v
CHAPTER	
1 INTRODUCTION.....	1
2 PRELIMINARIES AND MOTIVATION.....	4
3 ASSUMPTIONS AND ASYMPTOTIC RESULTS.....	10
Assumptions.....	10
Estimating the Number of Cointegrating Vectors.....	13
Consistency of the Cointegrating Vector.....	16
Estimation with a Time Trend.....	17
4 RESPONSE SURFACE ANALYSIS.....	19
Literature Review.....	19
Proposed Method.....	21
5 SIMULATIONS.....	24
Previous Methods.....	24
Proposed Method.....	26
6 CONCLUSION.....	30
7 REFERENCES.....	39
APPENDIX	
A MATHEMATICAL APPENDIX.....	41

## LIST OF TABLES

Table		Page
1.	Performance of the Hansen Test with Alternative Normalizations ....	32
2.	Estimated Parameters from Equation 3 .....	33
3.	Simulation Results, No Autocorrelation .....	34
4.	Simulation Results, Autocorrelation in the $f_t$ .....	35
5.	Simulation Results, Autocorrelation in the $g_t$ .....	36
6.	Simulation Results, Non-AR(1) Process .....	37

## LIST OF FIGURES

Figure	Page
1. Average Eigenvalues for the $P^{\text{th}}$ and $P+1^{\text{st}}$ Eigenvalues .....	45

## Introduction

Multiple economic time series display certain features that require specific techniques for parameter estimation. Classical multivariate techniques require stationary assumptions that restrict the behavior of the variables; when the variables appear non-stationary, alternative methods are required for testing procedures. On the other hand, cointegrated data series appear non-stationary while certain relationships between these variables appear stationary. This paper proposes estimating cointegrating relationships using a canonical correlation method that is easy to implement and provides consistent estimators of the span and number of cointegrating vectors.

Cointegrated series are individually nonstationary. However, certain linear combinations of nonstationary time series are stationary. Cointegration research dates back to Granger (1987) and Engle and Granger (1987). Engle and Granger (1987) use an error-correction representation which models current differences in the observed variables as linear functions of their lagged levels with, or without, lagged differences. The linear functions of the lagged levels are known as error-correction components. The stationarity of the differenced data and the inclusion of non-stationary lagged level as explanatory variables suggest certain linear combinations of the lagged level data are stationary.

Johansen (1988) proposes estimating the cointegrating relationships using a full information maximum likelihood method. Using a finite order vector autoregression with independent and identically distributed Gaussian error terms, Johansen (1988) provides a consistent estimator for the number of cointegrating

relationships. Here, likelihood ratio methods are used to estimate the number of cointegrating vectors, coefficients in the vectors and restrictions on the vectors. All of these results are derived under the assumption that level data follow a vector autoregressive (VAR) process of a finite order.

Likelihood methods may provide misleading results when the data-generating process is mis-specified. For example, selecting the correct number of lags to use is important in the estimation. If too few lags are used, the error term will be autocorrelated. Also, the level data may follow more general vector autoregressive and moving average (VARMA) processes that cannot be written as VARs of finite orders. Here, the error vector will follow a VMA(q) process. In both instances, the error terms will not be identically and independently distributed over time as the model assumes.

Even if level data follow a VAR process, Johansen's estimation and testing procedures may fail to produce reliable inferences in finite samples. Toda (1995) investigates the performance of the test to estimate the number of cointegrating vectors when there are two variables observed over 100 time periods with varying specification for the data generating process of the error term. He finds that the testing procedure for the number of cointegrating vectors suffers from a lack of power when there is a high degree of autocorrelation in the error terms; with a high degree of autocorrelation, the testing procedure says there is no cointegration when in fact, one cointegration does exist in the data.

In addition to poor performance when there is a large amount of autocorrelation in the error term, the testing procedure has poor results when the

data generating process for the error term is mis-specified. For example, simulations reported below show that the testing procedure does not perform well when the error term follows an MA(1) process. An MA(1) process has an AR( $\infty$ ) representation and using a finite number of lags, as assumed in Johansen (1986), is an approximation of the data series; however an MA(1) process can be approximated using an AR(d) process with sufficient d, yet the choice of p further complicates the estimation procedure.

Furthermore, Cheung and Lai (1995) show that the performance of the estimation procedure begins to deteriorate as the number of variables increases. The Johansen's (1986) testing method selects too few cointegrating vectors in the data. Further, simulations below show that this problem is exacerbated when the autocorrelation in the error term increases. These results extend the results in Toda (1995) to situations with more observed variables.

Other methods involving more general data-generating processes for the innovation terms have been developed. Phillips (1991) develops a regression framework for estimating cointegrating vectors with a general data-generating process. In this setup, one of the observed variables is treated as a response variable while the remaining variables are explanatory variables. Using a triangular system, Phillips (1991) produces a consistent estimator for the number of cointegrating vectors. The estimation procedure requires one of the values of the cointegrating vector normalized to be unity. However, simulations below show the choice of normalization is crucial for estimation, and an improper normalization can provide misleading results.



The estimation procedure in this paper involves forming canonical correlations between the differences in the observed variables and the lagged levels of the observed variables. Solutions to this problem involve eigenvectors of a matrix involving the variance of and the covariance between the levels and differences. Theoretical convergence rates of the eigenvalues similar to Bai and Ng (2002) are used in order to estimate the number of cointegration vectors as well as their span. This estimation procedure allows for general data-generating processes for the innovation terms and produces estimators which are consistent.

Our method is similar to Johansen's in the sense that both methods use canonical correlation method. We estimate cointegration vectors and the number of cointegration vectors analyzing unconditional canonical correlations between current values of differenced variables and one-period lagged level variables. The Johansen (1986) method uses canonical correlations between current values of differenced variables and one-period lagged level variables conditional on lagged values of differenced variables. Also, both methods use asymptotic methods to derive the results.

However, there are a couple of differences between the estimation methods in Johansen (1986) and our procedure. First, Johansen (1986) uses maximum likelihood estimation and assumes a specific functional form for the differenced data. On the other hand, we do not assume a specific functional form but instead make assumptions that ensure partial sums constructed from the error terms converge to functions of Brownian motion. Second, estimating the number of cointegrating vectors in Johansen (1986) is done using a distribution which

does not incorporate certain time-series features in data. However, our procedure for estimating the number of cointegrating vectors incorporates time-series properties in the data and adjusts accordingly.

### Preliminaries and Motivation

We begin by defining the data-generating process for the data. There are  $(N+P)$  observed variables over  $T$  time periods. Let  $z_{it}$  be the observed variable  $i$  ( $= 1, 2, \dots, N+P$ ) at time  $t$  ( $= 1, \dots, T$ ). The  $z_{it}$  are assumed all  $I(1)$  variables. Define an  $N \times 1$  vector of  $I(1)$  variables  $u_t = (u_{1t}, u_{2t}, \dots, u_{Nt})'$ . The variables in  $u_t$  follow a multivariate unit root process:

$$u_t = \sum_{s=1}^t g_s, \quad (1)$$

where  $g_t = (g_{1t}, g_{2t}, \dots, g_{Nt})'$  is an  $N \times 1$  vector of  $I(0)$  variables with  $E(g_{is}) = 0$ ,  $\forall i, s$ . Further, define a  $P \times 1$  vector of  $I(1)$  variables  $f_t = (f_{1t}, \dots, f_{Pt})'$ . The  $(N+P) \times 1$  vector of observed variables  $z_t = (z_{1t}, z_{2t}, \dots, z_{N+P,t})'$  is generated by the  $f_t$  and  $u_t$  variables. The variables  $f_t$ ,  $g_t$  and  $u_t$  are unobserved.

Using the above notation, the data-generating process for the observed variables is described by

$$z_t = Bf_t + Cu_t \quad (2)$$

Here,  $B$  is an  $(N+P) \times P$  matrix of the loadings on the  $I(0)$  variables in  $f_t$  and  $C$  is an  $(N+P) \times N$  matrix of the loadings on the  $I(1)$  variables in  $u_t$ . The combined

matrix  $[B, C]$  is of full column rank. Using the notation in (2), we can describe the complete data in matrix form by

$$Z = FB + UC \quad (3)$$

where  $Z = (z_1, z_2, \dots, z_T)'$ ,  $F = (f_1, f_2, \dots, f_T)'$ , and  $U = (u_1, u_2, \dots, u_T)'$ .

Using the above equations, the observed  $Z$  variables are linear combinations of the unobserved  $F$  and  $U$  variables. The non-stationarity in  $Z$  comes from the non-stationarity in the  $U$  variables. Removing the effect of the  $U$  variables from the  $Z$  variables leaves only the  $I(0)$  variables. Granger (1987) and Granger and Weiss (1983) define the components of an  $I(1)$  vector, say  $x$ , as being cointegrated of order one, if there exists a non-zero vector  $\alpha$  such that  $\alpha'x = I(0)$ . If we choose an  $(N + P) \times P$  matrix orthogonal to  $C$  (that is,  $A'C = 0_{(N+P) \times P}$ ), we have

$$Az_t = ABf_t = I(0) \quad (4)$$

Thus, each column of  $A$  is a cointegrating vector for  $z_t$ . Estimating the dimension and span of  $A$  is the primary focus of this paper.

The model in (2) can be generalized to include a linear time trend and non-zero intercept. A time trend can be introduced by specifying  $\tilde{g}_t = g_t + \delta t$  where  $\delta$  is an  $N \times 1$  vector. A non-zero intercept can be introduced to equation (2) by specifying  $\tilde{f}_t = f_t + \gamma$ , where  $\gamma$  is a nonzero  $P \times 1$  vector. With these additions, the model in (2) can be written as

$$z_t = Bf_t + Cu_t = B\tilde{f}_t + C\tilde{u}_t + B\delta + tC \quad (5)$$

It should be noted that the time trend component is in the space of  $\mathcal{C}$ ; this specification restricts the number of time trends to be no more than the rank of  $\mathcal{C}$ . As well, when the time trend is in the space of  $\mathcal{C}$ , the columns of  $A$  can still be interpreted as cointegrating vectors. However, with the inclusion of the intercept term, some of the cointegrating relationships may have non-zero expected value.

Estimating the number of cointegrating relations in (2) is achieved by maximizing the sample canonical correlations between the differenced data ( $\Delta z_t = z_t - z_{t-1}$ ) and the lagged levels ( $z_{t-1}$ ). The solution to this problem has an eigenvector representation; further, the estimated eigenvalues are the squared canonical correlations between the  $\Delta z_t$  and  $z_{t-1}$ . It is shown in the next section that if there are  $P$  cointegrating relations, there will be  $P$  eigenvalues which are  $O_p(1)$ . The remaining  $N$  eigenvalues will be  $O_p(T^{-1})$ . The divergence in the order of the eigenvalues forms the basis for the testing procedure.

There are two approaches for determining the number of eigenvalues which are  $O_p(1)$ . The first involves constructing a threshold level. The number of eigenvalues larger than this threshold value is an estimate of the number of  $O_p(1)$  eigenvalues. Bai and Ng (2002) utilize a threshold level in estimating the number of factors in stationary data with a large number of cross-sectional units and times series observations. The second approach follows Ahn and Horenstein (2011). There, the eigenvalues are first sorted, and then eigenvalue ratios are constructed using the consecutive eigenvalues. The number of cointegrating relations is chosen as the largest of these eigenvalue ratios; thus, creating the

ratios eliminates the need for a threshold value. However, if the eigenvalues have different orders of convergence, the ratio test can be misleading. Further discussion of the merits of the ratio test can be found in Ahn and Horenstein (2011).

There exists previous research which estimates the number of cointegrating relations. Johansen (1988) uses maximum-likelihood methods. In the Johansen (1988) model, the differenced data are assumed to follow a Gaussian VAR( $d$ ) model, which is equivalent to

$$\Delta Z_t = \Pi_0 Z_{t-1} + \Pi_1 \Delta Z_{t-1} + \dots + \Pi_p \Delta Z_{t-d} + \varepsilon_t, \quad (6)$$

where  $\varepsilon_t$  are *i.i.d.*  $N(0, \Omega)$ . Here, the differenced data is a function of the lagged levels, a finite number of lagged differences and an error vector  $\varepsilon_t$ .

Although the lagged levels are nonstationary, multiplying the lagged levels by a cointegrating vector creates a stationary series. Let  $\Pi_0 = BA'$  where  $B$  is an  $(N + P) \times P$  matrix. Here,  $A'z_{t-1}$  is a vector of error-correction terms which are  $I(0)$ . Testing for cointegration in equation (6) involves testing the rank of the matrix  $\Pi_0$ .

If we difference (2), we have

$$z_t = Bf_{t-1} + Bf_t + Cg_t = B(A'B)^{-1}A'z_{t-1} + e_t \quad (7)$$

where  $e_t = Bf_t + Cg_t$ . The second equality follows if we solve (2) at time  $t-1$  for  $f_{t-1}$  using the orthogonality of  $A$  and  $C$ . The representation in equation (7) displays the differenced data as a function of the lagged levels plus error terms.

Comparing the models for the differenced data in (7) and (6), we can see that the matrix  $B(A B)A$  is equal to  $\Pi_0$  and  $\Pi_1\Delta z_{t-1} + \dots + \Pi_p\Delta z_{t-p} + \varepsilon_t$  is equal to  $e_t$ . If we assume that  $e_t$  follows an  $AR(k)$  process, we see that the model in Johansen (1986) is a special case of the model in (2). Suppose that the model in Johansen (1988) stipulates  $e_t$  can be represented using a finite number of lagged differences of  $z_t$  plus an error vector  $\varepsilon_t$ . However, there do exist some specifications in which this is not a proper assumption; to wit, if  $e_t$  follows a MA(1) process, the  $AR(k)$  assumption is not appropriate.

At the moment, we place no such restrictions on the  $e_t$  other than it has a moving average representation of finite or infinite order. Further, Johansen (1988) stipulates the covariance matrix of  $\varepsilon_t$  is not time variant; this precludes various GARCH models. The model in this paper allows the short-run covariance matrix to vary; however, the long run covariance matrix of  $e_t$  is assumed to converge to a fixed matrix.

In addition, Phillips and Ouliaris (1990) provides an alternative procedure to that in Johansen (1988). In this alternative procedure, the data has a triangular representation

$$z_t^1 = Wz_t^2 + \varepsilon_t^1 = Wz_{t-1}^2 + \varepsilon_t^1 + W \varepsilon_t^2 \quad (8a)$$

$$z_t^2 = z_{t-1}^2 + \varepsilon_t^2. \quad (8b)$$

Here,  $z_t^1$  and  $z_t^2$  are the observed variables where  $z_t^1$  and  $\varepsilon_t^1$  are  $P \times 1$ ,  $z_t^2$  and  $\varepsilon_t^2$  are  $N \times 1$ , and  $W$  is a  $P \times N$  matrix.

The testing procedure for cointegration involves regressing the  $z_t^1$  variables on the  $z_t^2$  variables and using the residuals to test for cointegration. Phillips and Ouliaris (1990) assume the  $\varepsilon_t^1$  and  $\varepsilon_t^2$  are assumed strictly stationary and ergodic. With these assumptions, the OLS estimators of the coefficients of  $W$  converge to functions of Brownian motion. Testing the coefficients of the cointegrating vector is done using confidence intervals from simulated Brownian motion; testing for the presence of cointegration is done using the residuals from the OLS estimation.

In addition, the triangular system of (8a) and (8b) is a special case of (2). This is so because, when the variables in  $z_t$  are arranged in an appropriate way, the system becomes a nonsingular transformation of (2). To show this, let  $A = (A_1', A_2')'$ , where  $A_1$  is a  $N \times N$  matrix. Assume momentarily that the variables in  $z_t$  are arranged so that  $A_1$  is invertible. Let  $W = -(A_1')^{-1} A_2'$ ; and

$$D = \begin{pmatrix} I_P & 0_{P \times N} \\ -W' & I_N \end{pmatrix} \quad (9)$$

Notice that  $D$  is a nonsingular matrix by construction. If we pre-multiply equation (2) by  $D$ , we obtain (8a) and (8b) with  $\varepsilon_t^1 = (I_P, -W)(Bf_t + Cu_t) = (I_P, -W)Bf_t$  and  $\varepsilon_t^2 = (0_{N \times P}, I_P)(B\Delta f_t + Cg_t)$ .

However, it is important to note that this result requires the matrix  $A_1$  to be invertible. When the variables in  $z_t$  are inappropriately arranged and  $A_1$  is not invertible, the triangular system of (8a) and (8b) becomes a misspecified model.

Finally, Bossaerts (1988) avoids the normalization issues above and estimates the number of cointegrating relations using canonical correlation methods. His procedure involves two steps. First, the canonical correlations between the current level of the observed variables and the lagged level of the observed variables are estimated. Canonical correlations are found by forming linear combinations of the current and lagged levels which maximize the correlation between the two; the canonical variables are the resulting linear combinations of the variables formed by this process.

Next, each canonical variable is regressed on its lagged level. The autoregressive coefficient from this regression is tested using unit root asymptotics developed in Phillips (1987); the number of regressions which do not have a unit root correspond to the number of cointegrating relations. However, the setup in Bossaerts (1988) uses the Phillips (1987) unit root tests to test for cointegration which has low power against stationary alternatives as documented in Phillips and Perron (1988) and Kwiatkowski, Phillips, Schmidt and Shin (1992).

## Assumptions and Asymptotic Results

### Assumptions

The following assumptions are from Phillips and Durlauf (1986) and describe the asymptotic behavior of partial sums constructed from the  $g_t$  and  $f_t$ . Define  $v_t = (f_t', g_t')$  as the stacked vector of innovations. These partial sums are



constructed from the  $(N + P) \times 1$  vectors  $\mathbf{s}_t = \sum_{q=1}^t \mathbf{v}_q$ . The following assumptions ensure these partial sums will converge to functions of Brownian motion. When  $\mathbf{v}_t$  is a covariance stationary process, the  $\mathbf{s}_t$  will converge to functions of Brownian motion; however, less restrictive assumptions allow for more general data-generating processes.

**Assumption A:**  $E(\mathbf{v}_t) = \mathbf{0}_{(N+P) \times 1}$ , for all  $t$ .

**Assumption B:**  $E(T^{-1} \mathbf{s}_T \mathbf{s}_T') \rightarrow \Sigma$ , where  $\Sigma$  is a positive definite matrix, as  $T \rightarrow \infty$ .

**Assumption C:**  $\{v_{it}^2\}$  are uniformly integrable, for all  $i = 1, 2, \dots, N + P$ .

**Assumption D:**  $\sup_t E(|v_{it}|^\beta) < \infty$  for some  $2 \leq \beta < \infty$  and all  $i = 1, 2, \dots, N + P$ .

**Assumption E:**  $E[T^{-1}(\mathbf{s}_{k+T} - \mathbf{s}_T)(\mathbf{s}_{k+T} - \mathbf{s}_T)'] \rightarrow \Sigma$ , as  $\min(k, T) \rightarrow \infty$ .

**Assumption F:** Either  $\varphi_M$  is of size  $-\beta / (2\beta - 2)$  or  $\beta > 2$  and  $\alpha_m$  is of size  $-\beta / (\beta - 2)$ .<sup>1</sup>

---

<sup>1</sup> For  $\sigma$ -algebras  $F$  and  $G$ ,

$$(F, G) = \sup_{\{F, G, P(\cdot) > 0\}} |P(\cdot | F) - P(\cdot | G)|;$$

$$\alpha(F, G) = \sup_{\{F, G\}} |P(\cdot, F) - P(\cdot, G)|.$$

Further, define the  $\alpha$ -algebras generated by  $\{v_a, \dots, v_b\}$  as  $F_a^b$  and the  $\sigma$ -algebras generated by  $\{s_a - s_b, \forall a \leq b\}$  as  $L_a^b$ . The measures  $\varphi_m$  and  $\alpha_m$  are defined as

$$\varphi_m = \sup_n \sup_{j \geq n+m} \varphi(F_1^n, L_{n+m}^j);$$

$$\alpha_m = \sup_n \sup_{j \geq n+m} \alpha(F_1^n, L_{n+m}^j).$$

Assumption G:  $T^{-1} \sum_{s=1}^T f_s f_s' \rightarrow_p H_1$  and  $T^{-1} \sum_{s=1}^T [(Bf_{s+1}', Cg_{s+1}')' Bf_s] f_s' \rightarrow_p H_2$  as  $T \rightarrow \infty$ , where  $H_1$  and  $H_2$  are both fixed matrices where  $\text{rank}(H_1) = \text{rank}(H_2) = P$ .

Assumptions A – F provide sufficient conditions for which partial sums constructed from the  $u_t$  converge to functions of Brownian motion. This is an asymptotic invariance result from Phillips and Durlauf (1986) which is used throughout the paper.<sup>2</sup> The term asymptotic invariance is used because regardless of the short-term dynamics of the process  $u_t$ , only the long-term covariance matrix  $\Sigma$  determines the shape of the Brownian motion.

Assumption B allows the  $v_t$  to have a moderate amount of cross-sectional heteroskedasticity. In addition, Assumption B also permits some serial dependence in the  $v_t$ . The influence of any one individual short-run covariance matrix is asymptotically negligible as the sample size increases. Only the long-term covariance matrix determines the shape of the limiting function of Brownian motion.

---

<sup>2</sup> Asymptotic invariance pertains to the limiting distribution of the sum

$$X_T(r) = \Sigma^{-1/2} T^{-1/2} \sum_{q=1}^{t^*} v_q,$$

where  $t^* = \max\{t \mid t \leq rT, r \in [0,1]\}$ . For two processes  $v_1^t$  and  $v_2^t$  with identical long-term covariance  $\Sigma$  but different short run dynamics,  $X_T(t) \rightarrow W(r)$  for both processes where  $W(r)$  is Brownian motion.

Assumptions C and D are technical assumptions which provide limiting results relating to the probability of outliers. If  $f_{it}$  and  $g_{it}$  are random variables with finite fourth moments, then Assumptions C and Assumption D hold. These assumptions are necessary for applying the functional central limit theorem. The parameter  $\beta$  regulates the probability of outliers and the persistence of these outlying observations. GARCH processes allow for time-varying, persistent variance that can permit significant outliers requiring a larger value for  $\beta$ ; Assumption F requires that the influence of these observations decay more quickly. For a more detailed discussion see Andrews (1988) and Li and Terasvitra (1999).

Assumptions E and F pertain to the limiting temporal properties of the distribution. Assumption E concerns the relationship between subsets of the unit root processes and the entire process  $\{u_{it}\}$ ; in the limit, any unit root process constructed from contiguous values of  $f_{it}$  and  $g_{it}$  will have a long-run covariance matrix equal to  $\Sigma$ . Assumption F limits the amount of serial dependence in the error terms. The amount of serial dependence in the  $v_t$  must become asymptotically negligible at the specified rates. Alternatively, the joint distribution of the error terms must become asymptotically independent at a specified rate, as the time between the error terms increases. For a detailed discussion see the Appendix in Phillips and Durlauf (1986).

Assumption G states the sample covariance matrix of the  $I(0)$  components converge in probability to a matrix that has rank  $P$  as  $T \rightarrow \infty$ .

Equation (7) shows  $z_t$  is a function of  $f_t$ ,  $f_{t+1}$  and  $g_{t+1}$ . The second part of Assumption G states that the sample covariance matrix between  $z_t$  and  $f_t$  must converge in probability to a matrix that has rank  $P$ .

With assumptions A – F, the limiting distributions of certain matrices are functions of Brownian motion. However, as mentioned above, these are only sufficient conditions. The only requirement is that the canonical correlations in the data approach random or fixed matrices at specified rates. It is the divergence in the speed of these canonical correlations that allows canonical correlations to be used as a means of estimating cointegrating relations.

### Estimating the Number of Cointegrating Vectors

The first  $k$  largest canonical correlations between the  $z_{t+1}$  and  $z_t$  are calculated by finding linear combinations of the variables which maximize the correlation. The problem can be stated as

$$\max_{\{\hat{A}, \hat{D}\}} Tr(\hat{A}' Z' Z \hat{D} / T) \quad (10)$$

subject to the constraints on the vectors such that  $\hat{A}' Z' Z \hat{A} / T = I_k$  and  $\hat{D}' \Delta Z' \Delta Z \hat{D} / T = I_k$ . Here,  $\Delta Z = (\Delta Z_2, \dots, \Delta Z_{T+1})'$ ,  $\hat{A}$  and  $\hat{D}$  are  $(N + P) \times k$  matrices, and  $I$  is  $k \times k$  identity matrix, and  $Tr$  is the trace operator. The constraints in the maximization problem require the canonical variables  $Z \hat{A}$  and  $\Delta Z \hat{D}$  to have unit variances and zero covariances; the correlations between the  $Z \hat{A}$  and  $\Delta Z \hat{D}$  are the squared canonical correlations.

Setting up the Lagrangian and solving the constrained maximization, we have the following first order conditions

$$\hat{\Omega}_{\Delta Z, \Delta Z}^{-1} \hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z} \hat{D} = \hat{D} \hat{\Xi} \quad (11)$$

Here,  $\hat{\Omega}_{X, Y} = X'Y / T$  for any matrices  $X$  and  $Y$ , and  $\hat{\Xi}$  is an  $k \times k$  diagonal matrix with the squared Lagrangian multipliers along the diagonal,  $\hat{\lambda}_j$ . The solution  $\hat{D}$  contains the eigenvectors of the matrix  $\hat{\Omega}_{\Delta Z, \Delta Z}^{-1} \hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z}$ ; the Lagrangian multipliers are the associated squared eigenvalues. From the constraint placed on the problem, all of the eigenvalues are between zero and one.

**Theorem 1:** With assumptions A – G and when  $P \geq 1$ , the  $P$  largest eigenvalues are strictly positive and  $O_p(1)$ , while the  $N$  smallest eigenvalues are  $O_p(T^{-1})$ .

With assumptions A – G and when  $P = 0$ , all of the eigenvalues are  $O_p(T^{-1})$ .

Theorem 1 states that the number of eigenvalues in the data approach zero at different rates will depend on the number of cointegrating relations. Alternatively, the number of eigenvalues which are bounded away from zero is equal to the rank of  $A$ . Further, because the eigenvalues are also squared correlations, the eigenvalues are all bounded above by unity. Similarly, the smallest  $N$  eigenvalues will approach zero at speed  $T^{-1}$ . This divergence in the eigenvalues provides the basis for testing the number of cointegrations.

Selecting the true number of  $p$  is related to previous research on model selection which trade-off model performance and a penalty for overfitting. Cragg and Donald (1997) estimate the rank of a matrix by minimizing an objective function where the penalty function increasing in the choice of the matrix rank. Bai and Ng (2002) estimate the number of  $I(0)$  factors where the penalty function is designed to approach zero at a specified convergence rate as both the cross-sectional and time dimension become large. Moon and Perron (2007) use a modified estimator similar to that of Bai and Ng (2002) to estimate the number of non-stationary factors when  $N$  and  $T$  are large. Here, we are interested in creating an objective function which will consistently estimate the true number of cointegrations when  $N$  is small and  $T$  is large. Alternatively, Stock and Watson (1989) choose the number of factors to include by minimizing the mean squared forecast error of forecasts in out of sample testing.

We propose estimating the number of cointegrations using a threshold function. The threshold function is designed to approach zero at a rate slower than  $T^{-1}$ ; in this way, it places an upper bound on the eigenvalues which are not formed from cointegrating relations. Any threshold function which is  $O_p(T^{-\alpha})$  for  $\alpha \in (0,1)$  will correctly select the number of cointegrating relations in the data.

For any threshold function  $f(T)$  which is  $O_p(T^{-\alpha})$ , define  $\hat{P} = \max\{j | \hat{\lambda}_j > f(T)\}$ . Here,  $\hat{P}$  is the estimated number of cointegrating relations estimated as the number of eigenvalues which are above the value of the

threshold function for a given  $T$ . Alternatively, the threshold function can be included in an objective function as a penalty function similar to the related works above. For example, selecting the number of cointegrations can be formulated as  $\hat{P}^* = \arg \max_j \{\sum_{s=1}^j (\hat{\lambda}_s - f(T))\}$ , where  $f(T)$  acts as a penalty function for over selecting the number of cointegrations.

Asymptotically, the functional form of  $f(T)$  and the resulting  $\alpha$  are flexible so long as  $f(T)$  is  $O_p(T^{-\alpha})$ . This flexibility allows the threshold function to incorporate features of the data in order to more accurately estimate the number of cointegrating relations in finite samples.

**Proposition 1:** With assumptions A – F,  $P \geq 0$  and for any  $f(T)$  which is  $O_p(T^{-\alpha})$ ,  $\lim_{T \rightarrow \infty} \Pr(\hat{P} = P)$ , where  $\hat{P} = \max \{j \mid \hat{\lambda}_j > f(T)\}$ .

Proposition 1 states that using the eigenvalues in conjunction with a particular form of threshold function provides a consistent estimator of the true number of cointegrating relations. With a nonstationary series and no cointegrating relations, this implies we have  $(N + P)$  unit root processes. Various functional forms for the threshold function which exploit information in the data sample are explored below.

### Consistency of the Cointegrating Vector

Having shown using a threshold function to consistently estimate the number of cointegrating relations, the next step is to show the eigenvectors

associated with these  $P$  largest eigenvalues lie in the span of the true  $A$ . Define  $\hat{A}_p$  and  $\hat{D}_p$  as the eigenvectors associated with the  $P$  largest eigenvalues in the data from (9) associated with the  $\hat{P}$  largest eigenvalues. The estimators  $\hat{A}_p$  and  $\hat{D}_p$  are related through

$$\hat{A}_p = (Z'Z)^{-1}(Z'\Delta Z)\hat{D}_p\hat{\Xi}^{1/2}. \quad (12)$$

The term on the right hand side is the OLS estimator obtained by regressing the differenced data on the lagged levels multiplied by the estimator  $\hat{D}_p\hat{\Xi}^{1/2}$ . OLS estimates in cointegrated series have been studied at length in Phillips and Durlauf (1986), Phillips and Ouliaris (1990) and Phillips and Hansen (1990). As mentioned above, these regressions all require normalization where one variable is regressed on the remaining variables. However, when the data is generated by (2), the eigenvectors from canonical correlation analysis consistently estimate the space of the true  $A$ .

**Theorem 2:** With assumption A – G and  $P \geq 1$ ,  $\hat{A}_p = A(A'A)^{-1}\hat{D}_p\hat{\Xi} + o_p(T^{-1})$ .

Theorem 2 states that the vectors in  $\hat{A}_p$  are  $T$ -consistent estimators for the space of the true  $A$ .

Estimation with a Time Trend



Cointegrating relationship in the data can still be estimated when the data includes a time trend and the expected values of the cointegrating relationship is not zero; these correspond non-zero values for  $\gamma$  and  $\delta$  in (5). Estimating the time trend is not the focus in this paper; however, the intercept and time trend must be accounted for when estimating. Forming an alternative theory is avoidable by transforming the data before estimation. Previously, it was shown the largest  $P$  eigenvalues of the matrix  $\hat{\Sigma}_{Z,Z}^{-1} \hat{\Sigma}_{Z,Z}^{-1} \hat{\Sigma}_{Z,Z}$  are  $O_p(1)$  while the smallest  $N$  eigenvalues are  $O_p(T^{-1})$ . Transforming the data to remove the time trend produces a data series where the eigenvalues will have the same rates of convergence as in Theorem 1.

Transforming the data involves removing the intercept; removing the intercept is not necessary. This is done using the orthogonal complement of the matrix  $M(\tau) = I_T - \tau(\tau'\tau)^{-1}\tau'$ . Here,  $\tau$  is a vector with row ' $i$ ' entry equal to  $i$ ; in other words,  $\tau$  is a vector with a time trend. The appendix shows that multiplying the level data by this matrix removes the time trend while not affecting the orders of convergence of the eigenvalues; define the resulting data series as the *transformed data*. After transforming the data using  $M(\tau)$ , the estimated number of cointegrations consistently estimates the rank of  $A$ .

**Theorem 3:** With assumptions A – G, when the data is generated by equation (5), and when  $P \geq 1$  the  $P$  largest eigenvalues of the transformed data are strictly

positive and  $O_p(1)$ , while the  $N$  smallest eigenvalues are  $O_p(T^{-1})$ . With assumptions A – G and when  $P=0$  all of the eigenvalues are  $O_p(T^{-1})$ .

Theorem 3 states that the orders of convergence of the eigenvalues of the transformed data are identical to the orders of convergence of the eigenvalues of the data when a time trend is not present. Testing procedures for the number of cointegrating relationships when there is a time trend proceed in the same manner as when there are not a time trend.

Proposition 2: With assumptions A – G, when the data is generated by equation (5),  $P \geq 0$ , when the data contains a time trend and intercept and for any  $f(T)$  which is  $O_p(T^{-\alpha})$ ,  $\lim_{T \rightarrow \infty} \Pr(\hat{P} = P)$  where  $\hat{P} = \max\{j | \hat{\lambda}_j > f(T)\}$ .

## Response Surface Analysis

### Literature Review

Some previous studies have detailed poor finite sample properties of the ML estimator. For example, Cheung and Lai (1993) document the size of the MLE estimator for the number of cointegrating vectors increases, the number of variables increases and the data-generating process is mis-specified. The authors

find that when the number of variables ( $N + P$ ) increases the MLE overestimates the true number of cointegrations. In addition, when the innovation terms are finite moving-average processes, the MLE requires three or more lags for appropriately sized tests.

In addition to describing the performance of the MLE in finite samples, Cheung and Lai (1993) increase finite sample MLE performance by using response surface analysis which provides modified test statistics of the Johansen (1986) estimator for the number of cointegrations. Response surface analysis involves fitting response variables to control variables that are under the control of the researcher. The control variables under the control of the authors are the Ahn and Reinsel (1990) scaling factor,  $(T - (N + P)d)/T$ , where  $T$ ,  $N$  and  $P$  are as defined above and  $d$  is the number of lagged differences as in (6); the response variables are the ratios of the finite sample quantiles calculated from simulated data to their corresponding asymptotic critical values. This estimation procedure is designed to improve finite sample performance while adhering to asymptotic results.

Specifically, Cheung and Lai (1993) simulate data for various data configurations and calculate sample percentiles of the test statistics corresponding to specific confidence levels. Next, the ratios of these sample quantiles are divided by the asymptotic critical value of the test statistic. Finally, these ratios are regressed on an intercept and the Ahn Reinsel (1990) scaling factor,  $(T - (N + P)d)/T$  as defined above. The estimated slope and intercept

coefficients are found to improve the finite sample performance of the Johansen (1986) estimator.

Response surface analysis is also used in MacKinnon (1996) to produce functional forms for critical values and  $p$ -values of the MLE using simulated data. The critical value functions are affine functions of  $T^{-1}$  and  $T^{-2}$  with a non-zero intercept. As  $T$  approaches infinity, the finite sample corrections due to  $T^{-1}$  and  $T^{-2}$  approach zero. Therefore, the intercepts are interpreted as the asymptotic critical values and  $p$ -values from the simulated data.

Ho and Sorenson (1996) and Cheung and Lai (1995) investigate the number of lags to use in the error-correction model and finite-sample MLE performance when the number of variables and cointegrations is allowed to grow. In finite samples, the authors find the MLE estimator has low power when  $N + P > 2$ . In fact, the MLE tends to select too many cointegrations in the simulated data. As  $N + P$  grows, the MLE estimator has low power against hypothesis tests where the null assumes more cointegrations than the true number. Binder, Hsiao and Pesaran (2005) develop methods for estimating cointegrating relations with small  $T$  and large  $N$  that rely on  $N$  approaching infinity and  $T$  fixed. However, the purpose of this study is to investigate the relative performance of the competing estimators when  $N$  is small and  $T$  is large.

The methods developed in this paper do not use a parameter-free asymptotic distribution to test for the number of cointegrations. Instead, the estimator developed here has a flexibility in parameter choice. Any threshold function which is  $O_p(T^{-\alpha})$ , where  $0 < \alpha < 1$ , can be used to consistently estimate

the number of cointegrating relationships. The  $\alpha$  can be chosen such that the threshold function performs well in finite samples.

### Proposed Method

For the simulations here, the threshold function,  $Th(T, N, R, S)$ , is assumed to be of the form

$$Th(T, N + P, R, S) = kT^{-\alpha} (N + P)^\delta R^\omega S^\theta \quad (13)$$

Here  $T$  is the number of time periods,  $N+P$  is the number of variables observed over the time period and  $R$  and  $S$  are functions explained below. For  $0 < \alpha < 1$ , the threshold function is  $O_p(T^{-\alpha})$  when all other variables are held constant. This functional form is chosen because both sides can be transformed into a linear function by taking the log of each side; after this transformation, the parameters can be estimated using least-squares methods.

In equation (13),  $R$  is a function of a weighted average of autocorrelations. Toda (1994) documents that the autocorrelation present in the innovations plays a large role in the accuracy of the test. Because of this, the weighted average of the autocorrelations is included. This value is calculated in four steps. First, the matrix of level observations is multiplied by the  $(N + P) \times (N + P)$  matrix of eigenvectors from (11) producing a  $T \times N$  matrix of candidate stationary components  $\hat{F} = Z\hat{A}$ . The variables in the matrix  $\hat{F}$  are candidate stationary components because only  $P$  of the  $(N + P)$  variables are stationary when the true

number of cointegrations is  $P$ . The  $(t, l)^{th}$  element of  $\hat{F}$  is  $\hat{f}_{it}$ . Second, the correlation between each  $\hat{f}_{it}$  and  $\hat{f}_{it-1}$ ,  $\hat{\rho}_i$ , is computed. These correlations measure the autocorrelation present in the candidate stationary components. Because there are only  $P$  stationary components in the data,  $N$  correlations will approach unity as the sample size increases.

Third, weights are calculated using the  $\hat{\rho}_i$ . For each component, the weight is calculated as

$$\hat{w}_i = \frac{(1 - |\hat{\rho}_i|)}{\sum_{i=1}^{N+P} (1 - |\hat{\rho}_i|)} \quad (14)$$

With this functional form, the  $\hat{f}_{it}$  with less autocorrelation will have greater weight; if  $R$  is computed using an equally-weighted average,  $R$  will increase as the number of non-stationary components increases. With the weights chosen using (2), the near unity autocorrelations from the non-stationary components are negligible. Finally,  $R$  is computed using<sup>3</sup>

$$R = 1 + \sum_{i=1}^{N+P} \hat{w}_i |\hat{\rho}_i|. \quad (15)$$

The weighed sum is added to unity in (15) to control for the situation when there is no autocorrelation in the stationary processes. When there is no autocorrelation

---

<sup>3</sup> Additional specifications for  $R$  were also considered. Using

$R = 1 + (\sum_{j=1}^{N+P} j^{-1})^{-1} \sum_{j=1}^{N+P} j^{-1} |\hat{\rho}_j|$ ,  $R = 1 + (\sum_{i=1}^{N+P} \exp(-j))^{-1} \sum_{i=1}^{N+P} \exp(-j) |\hat{\rho}_j|$ , and  $R = 1 + |\hat{\rho}_1|$ , where  $|\hat{\rho}_j|$  is the  $j^{th}$  largest value of all  $|\hat{\rho}_i|$  ( $i = 1, \dots, N+P$ ) and  $\hat{\rho}_1$  is the largest value of all  $\hat{\rho}_i$  ( $i = 1, \dots, N+P$ ). All provided similar results, but the weights in (2) provided an  $R$  that was the most accurate.

in the stationary processes,  $P$  of the  $\hat{\rho}_i$  will be close to zero and the weighted sum  $\sum_{i=1}^{N+P} \hat{w}_i |\hat{\rho}_i|$  will be close to zero. In order to prevent the threshold function from approaching zero as a result of low autocorrelation, the weighted sum is added to unity.

The variable  $S$  is an equally-weighted average of  $R^2$  when regressing  $z_{it+1}$  on  $z_t$  and a constant for all  $i=1, \dots, N+P$ . The  $R^2$  measures the amount of variation in the differenced data that the lagged levels can. When the  $R^2$  is small, the lagged levels do not explain a large amount of the variation in the differenced data. In this instance, the correlation between the differenced data and the lagged levels will be low and the resulting canonical correlations will be lower as well. Controlling for the average  $R^2$  in the data allows the threshold function to adjust accordingly.

Each of the variables in the threshold function influences the behavior of the  $P^{\text{th}}$  and  $(P+1)^{\text{st}}$  eigenvalues. Figure 1 shows the average 1<sup>st</sup> and 2<sup>nd</sup> largest eigenvalues for various values of  $N$  when  $P = 1$ ; in these simulations  $f_t$  and  $g_t$  follow an AR(1) process. Both eigenvalues are increasing with the number of variables. Further, the eigenvalues decrease as the serial correlation in the  $f_{it}$  and  $g_{it}$  increases; in fact, when  $\rho = 0.9$ , the average  $(P+1)^{\text{st}}$  eigenvalue is above the average  $P^{\text{th}}$  eigenvalue when  $N > 5$ . These and other simulations suggest the threshold function should take into account the number of variables as well as an estimate of the serial correlation in the innovations.

The choice of the parameters  $k$ ,  $\delta$ ,  $\psi$  and  $\theta$  determine the intercept of the threshold function while the parameter  $\alpha$  determines the speed at which the threshold function approaches zero. Ideally, the value of the threshold function will lie between the  $P^{\text{th}}$  eigenvalue,  $\hat{\lambda}_p$ , and the  $(P+1)^{\text{th}}$  eigenvalue,  $\hat{\lambda}_{p+1}$ . This suggests estimating the following equation

$$w \hat{\lambda}_p + (1 - w) \hat{\lambda}_{p+1} = kT \quad N \quad R \quad S \quad (16)$$

Taking logs

$$\ln(\tilde{\lambda}_{w,P,P+1}) = \ln(k) + \ln(T) + \ln(N) + \ln(R) + \ln(S) \quad (17)$$

Here  $0 < w < 1$ . This methodology will provide a good and practical estimator of a specific linear combination of the eigenvalues. However, there is no guarantee the parameter estimates are the parameters which will select the true number of cointegrating relationships with the greatest accuracy. Still, simulations will show that this methodology can provide parameter estimates which provide accurate estimates for the true number of cointegrating relationships.

The choice variable  $T$  is  $O(T)$ , but the term  $\hat{\lambda}_{w,P,P+1}$  is  $O_p(1)$ . Therefore, with finite  $N$  and  $R$ , the ordinary least-squares estimate for  $\alpha$  will be  $O_p(T^{-1})$ . The asymptotic behavior of  $T$  was not an issue in MacKinnon (1996) because the regression used  $T^{-1}$  as a regressor. Because of this, the estimation is done in two parts. First, a value for  $\alpha$  is chosen<sup>4</sup> and the parameters  $\delta$  and  $\varphi$  are estimated using ordinary least-squares with small values for  $T$  and various combinations of

---

<sup>4</sup> Here, possible  $\alpha$  are in the set  $\{0.01, 0.02, 0.03, \dots, 0.98, 0.99\}$



$N$ ,  $P$ , as well as the parameters governing the data generating process for the innovations. Here, the least-squares estimates  $\hat{\delta}(\alpha)$  and  $\hat{\psi}(\alpha)$  are functions of  $\alpha$ . In addition to storing the least-squares estimates, the percentage of times the threshold function is between the  $P^{\text{th}}$  and  $(P+1)^{\text{st}}$  eigenvalues,  $\pi(\alpha)$ , is stored as well. The estimated parameter values are the argmax for  $(\hat{\alpha})$ ,  $\hat{\alpha}$ , as well as the corresponding  $(\hat{\delta})$  and  $(\hat{\psi})$ .

## Simulations

### Previous Methods

The virtue of the proposed estimation procedure is that it does not require any specific normalization of the data. This flexibility precludes improper conclusions about the presence of cointegrating vectors. As mentioned above, the testing procedure in Phillips and Ouliaris (1990) does require this normalization. Table 1 displays results for simulations where there is only one cointegrating vector among 5 observed variables; here  $W$  is a  $1 \times 4$  row vector with a 1 in the first column and zeros in the remaining columns. This table shows that the Phillips and Ouliaris (1990) testing procedure correctly rejects the null of no cointegration in every simulation when one of the cointegrated variables is on the left-hand side of the regression. However, when both of the cointegrated variables are on the right-hand side of the regression, the testing procedure rejects the null of no cointegrations less than 4% of the time. These simulations show,

the method in Phillips and Ouliaris (1990) cannot detect any cointegrating relations which include only the variables in  $z_t^2$ .

The theory for determining the number of cointegrations uses asymptotic results as the number of time periods goes to infinity. However, because all data sets are finite, estimator performance in finite samples is of interest. Toda (1995) discusses the finite sample performance of the MLE estimator for the number of cointegrations developed in Johansen (1988). The experiment in Toda (1995) fixes the number of variables at  $N + P = 2$  and the number of time periods at  $T = 100$  while varying the temporal parameters governing the data-generating process for the innovations; in equation (2), the experiment is similar to varying the parameters describing the data-generating process for  $f_t$  and  $g_t$ .

The simulation procedure in Toda (1995) can be expressed using equation (2). When  $N = 2$ , there are no cointegrating vectors and the matrix  $C$  is set equal to an identity matrix. When  $N = 1$ ,  $A' = [1, 0]$  and  $C' = [0, 1]$ . The variables  $f_t$  and  $g_t$  each follow an AR(1) process. Toda (1995) investigates the performance of the Johansen (1986) estimator for the number of cointegrating vectors when the autocorrelation of the  $f_t$  and  $g_t$  is varied and the correlation between the innovations to the are varied  $f_t$  and  $g_t$ .

Toda (1995) concludes that for  $N + P = 2$ ,  $T = 100$  is too few observations for reliable statistics. He states that  $T > 300$  is required for good performance of the MLE uniformly over parameters governing the data-generating process. However, when the correlation between the innovations in

the non-stationary components and stationary components increases, the percentage of times the Johansen (1986) estimator correctly selects the number of cointegrating vectors increases. When the degree of autocorrelation in the innovation process increases, performance decreases.

### Proposed Method

The AR(1) model can capture many features present in data. Examining the error-correction model for the data can shed some light on the impact the serial correlation parameters have. The differenced data can be written as

$$\Delta Z_{it+1} = b'_i \Delta f_{t+1} + c'_i g_{t+1} = b'_i [(\rho_f - 1) f_t + h_t] + c'_i (\rho_g g_t + q_t) \quad (18)$$

As mentioned above, the cointegrating vectors are vectors which determine long-run economic relationships. In the model above, the cointegrating vectors are the  $(N + P) \times 1$  vectors in  $A$  and the deviations from these long-run relationships are the stationary components  $f_t$ . The serial correlation in  $f_t$  determines how long the variables remain away from their long-run relationships; the larger  $\rho_f$  is, the longer the stationary components will remain above zero, and the longer variables will remain away from their long-run level.

In addition, when the value of  $\rho_f$  is large, the correlation between the lagged stationary components and the differenced data decreases. From equation (18), the closer the value of  $\rho_f$  is to unity, the less the variance in the differenced data is explained by the lagged stationary components. As a result, the correlation between the lagged stationary components and the differenced data decreases.

Therefore, the eigenvalues associated with the canonical correlations will decrease as well.

As mentioned before, the parameters are estimated using simulated data with finite values of  $T$ . The innovations are generated using an autoregressive process AR(1) where  $f_{it} = \rho_f f_{it-1} + h_{it}$  and  $g_{jt} = \rho_g g_{jt-1} + q_{jt}$  where  $h_{it}$  and  $q_{jt}$  are iid  $N(0, 1 - \rho^2)$  for all  $i, j$  and  $t$ , and  $\rho = \rho_f$  or  $\rho_g$ ; the variance is chosen in order to normalize the long-run variance of the  $f_{it}$  and  $g_{jt}$ . The entries in the matrices  $A$  and  $C$  are all  $N(0, 1)$  random variables. The simulations are run using  $T = 150, 200, 250$ ;  $N = 2, 3, 4, 5$ ;  $P = 0, 1, 2, 3, 4$ , and  $\rho = 0, 0.1, 0.25, 0.5$ .

The estimated parameters are displayed in the first column of Table 2. The estimated value for  $\alpha$  is between zero and one. As the other variables are assumed to be finite, the threshold function will be  $O_p(T^{-0.23})$ . The threshold function will approach zero as the number of time series observations grows large while the largest  $P$  eigenvalues will be  $O_p(1)$ . Because of this, the threshold function can be thought of as a lower bound to the  $P$  largest eigenvalues and an upper bound to the  $N$  smallest eigenvalues.

Also, the value for  $\delta$  is between zero and one; because of this, the threshold function will grow without bound as  $N$  grows large. However, as  $N$  is assumed finite, the value of the threshold function in large samples will be determined by the number of time series observations. Further, the positive value of delta indicates the threshold function must be adjusted upwards as more

variables are observed. This upwards adjustment is a result of both the  $P^{th}$  eigenvalue decreasing with  $N$  and the  $(P+1)^{st}$  eigenvalue increasing with  $N$ .

The degree of autocorrelation in the  $f_{it}$  and  $g_{it}$  plays a role in determining the threshold value. The value for  $\psi$  is positive which indicates larger values of  $\rho$  require an upward adjustment of the threshold function. This result is intuitive. In the limit, as  $\rho$  approaches unity for the stationary components,  $f_{it}$ , the process  $f_{it}$  becomes  $I(1)$ . It must be noted that the latter estimated autocorrelations used in equation (2) will approach unity; however, the decreasing weights associated with the larger auto correlations will underweight the  $\hat{\rho}_i$  used to calculate  $R$ . Therefore, additional  $\hat{\rho}_i$  will not play a significant role in determining  $R$  when  $N$  is increased. To this extent, changes in  $N$  will not cause significant changes in  $R$ .

Three tests are compared when estimating the number of cointegrating relationships in the data. The three tests include Johansen (1986) likelihood method procedure, the Ahn and Reinsel (1988) scaled version of the Johansen (1986) testing procedure and the threshold method outlined above. The Ahn and Reinsel (1988) method adjusts the test statistics from Johansen (1986) by a factor of  $(T - (N + P)d)/T$  where  $d$  is the number of lagged differences included in the estimation and compares the scaled test statistic to the critical value. The scale factor is designed to improve the finite sample performance of the Johansen (1986) testing procedure. This scaling factor approaches 1 as  $T$  grow large and decreases as the number of variables increases.

Table 3 displays the performance of the procedures when the innovations in the data are iid  $N(0,1)$  random variables. The table shows that all tests have increasing performance when  $T$  grows. This is to be expected as all three procedures are based on large  $T$  asymptotics. However, the table also shows that the Johansen (1986) test and the Ahn and Reinsel (1988) test have poor performance when  $(N + P)$  increases. When  $N + P > 2$  and  $T = 100$  the Johansen (1986) method selects too few cointegrations. However, the threshold method performs well and selects the true number of cointegrations more than 99% of the time.

Table 4 shows the performance of the testing procedures when the  $f_{it}$  are autocorrelated. The autocorrelation coefficient is set equal to 0.5 in all situations. Simulations show that both the Johansen (1986) method and the Ahn and Reinsel (1988) method perform poorly in samples where there are only one cointegrating relation; this performance deteriorates when  $T$  is smaller. Further, in samples where  $T = 100$ , these procedures are biased downwards when selecting the number of cointegrations. However, the threshold has superior performance in these simulations. When  $T = 200$ , the threshold method selects the true number of cointegrating relations more than 98% of the time.

There can also be situation where there is autocorrelation in the  $g_{it}$ . Table 5 shows the simulation results for this situation. In this setting, all three methods perform well. However, in simulations where  $T = 100$ , each testing procedure has certain configurations of the data where it performs poorly. The Johansen (1986) and Ahn and Reinsel (1988) methods perform poorly when  $N + P = 4$  and  $P = 2$

or 3; alternatively, the threshold method performs poorly when  $N + P = 4$  and  $P = 0$  or 1.

Of course, the threshold function here is fit using data generated by an AR(1) process. As a robustness check, the performance of the threshold function is checked using data simulated using an alternative data generating process. First,  $(N+P) \times 1$  vectors  $v_t$  are generated where the  $v_t$  follow an AR(1) process with autocorrelation of 0.5 and iid error terms. Next, the  $f_{it}$  and  $g_{jt}$  are constructed from the  $v_t$  by using the equation  $f_t = M_t v_t$ . Here,  $M_t$  is a matrix of the first  $P$  rows of a  $(N+P) \times (N+P)$  random permutation matrix. With this construction, correlation between the  $f_{it}$  and  $f_{it-1}$  is random. Table 6 shows the performance of the threshold estimator and the other competing estimators when the  $f_{it}$  and  $g_{jt}$  follow this process; it is clear from the table that the threshold function outperforms all of the estimators.

## Conclusion

This paper builds on previous literature estimating cointegrating relations. The methods developed in this paper are based upon the asymptotic results of Phillips and Durlauf (1986) and permit flexible data-generating process for the innovations to the data series. The eigenvalue and eigenvector methods developed rely on the order of convergence of eigenvalues computed from the data. The resulting asymptotics of the data do not require exact specifications for the data-generating process but only assumptions on the limiting behavior of the

data-generating process. Previous research in Johansen (1988) has implemented full-information maximum likelihood methods which required Gaussian errors with an autoregressive specification. As shown above, there are examples for which the maximum likelihood methods are not consistent and do not have good finite sample properties. The proposed method is consistent over a wide variety of data-generating processes for the errors. However, simulations have shown the maximum likelihood methods do have excellent properties when the error process is correctly specified. Still, the proposed estimator consistently estimates the number of cointegrating vectors and the space of the cointegrating vectors for a large class of data generating processes.

In addition, this methodology can be used when there is a time trend in the data generating process. Detrending the data before estimating the canonical correlation removes the time trend but retains the order of convergence differential between the first  $p$  eigenvalues and latter eigenvalues. Even if there is no time trend present, de-trending the data in this manner will not affect any of the convergence results.

The focus in this paper has been on developing simple eigenvector and eigenvalue methods for estimating cointegrating relations in nonstationary data. Methods for consistent estimation of the size and space of the cointegrating relationships are developed. Simulations show the proposed estimator outperforms maximum likelihood methods when there is misspecification in the maximum-likelihood objective function and is on par when the model is correctly specified.



Table 1

*Performance of the Hansen Test with Alternative Normalizations*

Number of Additional Variables in the Regression	Index of the Variable Contained on the Left	Average Value of the $Z_p$ Test Statistic	Fraction of Times the Null of No Cointegration is Rejected	Average $Z_t$ Statistic	Fraction of Times the Null of No Cointegration is Rejected
0	1	-495.4	1.00	-22.5	1.00
1	1	-492.4	1.00	-22.5	1.00
2	1	-489.5	1.00	-22.6	1.00
3	1	-486.7	1.00	-22.6	1.00
4	1	-484.0	1.00	-22.7	1.00
0	2	-511.2	1.00	-22.3	1.00
1	2	-518.0	1.00	-22.2	1.00
2	2	-523.8	1.00	-22.1	1.00
3	2	-528.6	1.00	-22.0	1.00
4	2	-532.5	1.00	-22.0	1.00
1	3	-6.610	0.02	-1.55	0.02
2	3	-10.77	0.02	-2.12	0.02
3	3	-14.85	0.03	-2.57	0.02
4	3	-19.05	0.03	-2.96	0.03
2	4	-10.76	0.02	-2.13	0.02
3	4	-14.76	0.02	-2.56	0.02
4	4	-18.82	0.03	-2.94	0.03
3	5	-14.74	0.02	-2.56	0.02
4	5	-18.83	0.03	-2.94	0.03
4	6	-19.17	0.03	-2.97	0.03

*Notes:* Data is simulated using the Phillips (1991) triangular representation as in (8a) and (8b). Here  $z_t^1$  is a scalar variable and  $z_t^2$  is a  $4 \times 1$  vector. The matrix  $W$  is a  $1 \times 4$  vector with a 1 in the first entry and zeros elsewhere; this implies the first and second variables are cointegrated. All error terms in  $\epsilon_t^1$  and  $\epsilon_t^2$  are iid  $N(0,1)$  random variables. The table shows the percentage of times the null hypothesis of no cointegration is rejected using the methods described in Phillips (1991). The first and second variables are always used in the regression. The column ‘number of additional variables in the regression’ shows the number of variables included in the regression in addition to the first and second variables. The column ‘Index of the variable contained on the left’ shows what variable is placed on the left-hand side of the regression.

Table 2

*Estimated Parameters from Equation 3*

Parameter	Estimate
K	1.538
A	0.230
$\Delta$	0.252
$\Phi$	-1.020
$\Theta$	-2.504

*Notes:* The parameter estimates below correspond to the threshold function  $kT^\alpha N^\delta R^\psi S^\theta$ . The threshold function is designed to produce a value between the  $P^{th}$  and  $(P+1)^{st}$  eigenvalues given value for  $T$ ,  $N$  and  $R$ .  $T$  is the number of time periods,  $N$  is the number of variables observed,  $R$  is a weighted average of the estimated autocorrelations computed from the estimated stationary components and  $S$  is an equally weighted average of the  $R^2$  from a regression of the differenced data in the lagged levels. The parameters are estimated using ordinary least-squares. Data is simulated using  $T = 150, 200, 250$ ;  $N + P = 2, 3, 4, 5$ ;  $P = 0, 1, 2, 3$  and  $|\rho| \leq 0, 0.1, 0.25, 0.5$ . The dependent variable in the regression is  $w\lambda_p + (1-w)\lambda_{p+1}$  where  $w = 0.66$  and the explanatory variables include a constant,  $\ln(T)$ ,  $\ln(N+P)$ ,  $\ln(R)$  and  $\ln(S)$ .

Table 3

*Simulation Results, No Autocorrelation*

$T$	$N+P$	$P$	Sample Averages			Accuracy		
			Johansen	A-R	EV	Johansen	A-R	EV
100	2	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
100	3	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
100	4	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
100	2	1	1.06	1.05	1.00	94.4%	94.8%	100.0%
100	3	1	0.93	0.90	1.01	92.4%	90.2%	99.8%
100	4	1	0.60	0.52	1.05	59.8%	52.4%	98.8%
100	3	2	2.05	2.04	2.00	93.0%	93.0%	100.0%
100	4	2	1.67	1.52	2.01	71.4%	60.4%	99.6%
100	4	3	2.97	2.87	3.00	85.8%	79.0%	100.0%
150	2	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
150	3	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
150	4	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
150	2	1	1.06	1.05	1.00	94.0%	94.8%	100.0%
150	3	1	1.00	1.00	1.00	100.0%	100.0%	100.0%
150	4	1	1.00	1.00	1.00	99.8%	99.6%	100.0%
150	3	2	2.07	2.06	2.00	93.4%	93.8%	100.0%
150	4	2	2.01	2.01	2.00	99.4%	99.4%	100.0%
150	4	3	3.06	3.06	3.00	93.6%	94.0%	100.0%
200	2	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
200	3	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
200	4	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
200	2	1	1.04	1.04	1.00	95.8%	95.8%	100.0%
200	3	1	1.01	1.01	1.00	98.8%	99.0%	100.0%
200	4	1	1.00	1.00	1.00	100.0%	100.0%	100.0%
200	3	2	2.04	2.04	2.00	96.0%	96.2%	100.0%
200	4	2	2.00	2.00	2.00	99.8%	99.8%	100.0%
200	4	3	3.06	3.05	3.00	94.2%	94.6%	100.0%

*Notes:* Data is generated with various values of  $T$ ,  $N$  and  $P$  for 500 simulations.

The sample averages are the average values of the number of estimated cointegrations. The Accuracy is the number of times the true number of cointegrations is selected.

Table 4

*Simulation Results, Autocorrelation in the  $f_t$* 

$T$	$N+P$	$\rho_F$	$P$	Sample Averages			Accuracy		
				Johansen	A-R	EV	Johansen	A-R	EV
100	2	0.5	0	0.00	0.00	0.00	99.8%	99.8%	100.0%
100	3	0.5	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
100	4	0.5	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
100	2	0.5	1	0.82	0.80	0.98	73.8%	73.0%	98.2%
100	3	0.5	1	0.18	0.14	1.00	17.4%	13.8%	99.0%
100	4	0.5	1	0.04	0.02	1.06	3.8%	1.8%	98.0%
100	3	0.5	2	0.84	0.73	1.92	22.0%	16.0%	96.2%
100	4	0.5	2	0.11	0.07	1.93	0.8%	0.4%	96.4%
100	4	0.5	3	0.48	0.34	2.51	1.0%	0.6%	83.8%
150	2	0.5	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
150	3	0.5	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
150	4	0.5	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
150	2	0.5	1	1.04	1.04	1.00	95.6%	95.6%	100.0%
150	3	0.5	1	0.75	0.71	1.00	74.0%	70.0%	99.8%
150	4	0.5	1	0.28	0.22	1.00	28.4%	22.4%	99.8%
150	3	0.5	2	2.01	1.97	1.98	86.4%	83.4%	99.2%
150	4	0.5	2	0.86	0.74	1.96	22.4%	18.0%	98.0%
150	4	0.5	3	2.41	2.24	2.83	50.8%	43.8%	94.2%
200	2	0.5	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
200	3	0.5	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
200	4	0.5	0	0.00	0.00	0.00	100.0%	100.0%	100.0%
200	2	0.5	1	1.06	1.06	1.00	94.0%	94.2%	100.0%
200	3	0.5	1	1.00	1.00	1.00	97.8%	97.6%	100.0%
200	4	0.5	1	0.79	0.76	1.00	79.0%	75.8%	99.8%
200	3	0.5	2	2.06	2.06	2.00	94.4%	94.4%	100.0%
200	4	0.5	2	1.89	1.86	1.98	89.8%	87.6%	99.2%
200	4	0.5	3	3.07	3.05	2.95	92.6%	93.2%	98.2%

*Notes:* Data is generated with various values of  $T$ ,  $N$  and  $P$  for 500 simulations. The sample averages are the average values of the number of estimated cointegrations. The Accuracy is the number of times the true number of cointegrations is selected.

Table 5

*Simulation Results, Autocorrelation in the  $g_t$* 

$T$	$N+P$	$\rho_g$	$P$	Sample Averages			Accuracy		
				Johansen	A-R	EV	Johansen	A-R	EV
100	2	0.50	0	0.00	0.00	0.04	100.0%	100.0%	98.6%
100	3	0.50	0	0.00	0.00	0.26	100.0%	100.0%	93.4%
100	4	0.50	0	0.00	0.00	1.09	100.0%	100.0%	78.2%
100	2	0.50	1	1.06	1.06	1.00	94.0%	94.2%	100.0%
100	3	0.50	1	0.95	0.92	1.16	92.6%	90.2%	94.6%
100	4	0.50	1	0.67	0.58	1.85	66.8%	57.8%	78.8%
100	3	0.50	2	2.05	2.04	2.01	93.8%	94.6%	99.4%
100	4	0.50	2	1.71	1.58	2.13	75.2%	65.4%	95.8%
100	4	0.50	3	3.00	2.94	3.00	90.2%	86.4%	99.8%
150	2	0.50	0	0.00	0.00	0.01	100.0%	100.0%	99.8%
150	3	0.50	0	0.00	0.00	0.08	100.0%	100.0%	98.0%
150	4	0.50	0	0.00	0.00	0.47	100.0%	100.0%	90.6%
150	2	0.50	1	1.05	1.05	1.00	95.0%	95.2%	100.0%
150	3	0.50	1	1.01	1.01	1.02	99.2%	99.4%	99.4%
150	4	0.50	1	1.00	1.00	1.24	99.4%	99.6%	94.0%
150	3	0.50	2	2.07	2.07	2.00	92.6%	92.8%	100.0%
150	4	0.50	2	2.02	2.02	2.05	98.0%	98.2%	98.4%
150	4	0.50	3	3.06	3.05	3.00	94.2%	94.6%	100.0%
200	2	0.50	0	0.00	0.00	0.01	100.0%	100.0%	99.6%
200	3	0.50	0	0.00	0.00	0.02	100.0%	100.0%	99.4%
200	4	0.50	0	0.00	0.00	0.16	100.0%	100.0%	96.8%
200	2	0.50	1	1.07	1.06	1.00	93.2%	93.6%	100.0%
200	3	0.50	1	1.02	1.02	1.00	98.4%	98.4%	100.0%
200	4	0.50	1	1.00	1.00	1.02	100.0%	100.0%	99.4%
200	3	0.50	2	2.06	2.06	2.00	94.2%	94.4%	100.0%
200	4	0.50	2	2.01	2.01	2.01	99.4%	99.4%	99.8%
200	4	0.50	3	3.06	3.06	3.00	93.8%	94.4%	100.0%

*Notes:* Data is generated with various values of  $T$ ,  $N$  and  $P$  for 500 simulations. The sample averages are the average values of the number of estimated cointegrations. The Accuracy is the number of times the true number of cointegrations is selected.

Table 6

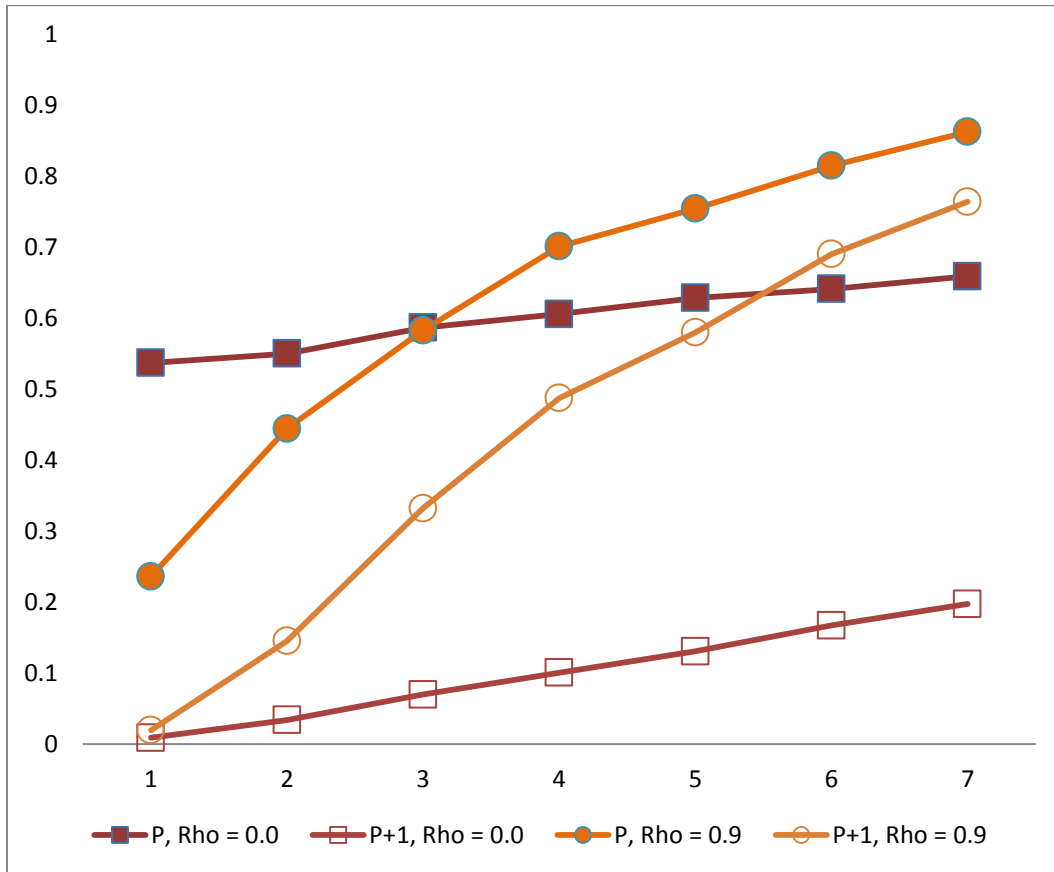
*Simulation Results, non-AR(1) Process*

$T$	$P$	$N+P$	$\rho_v$	Sample Averages			Sample Averages		
				Johansen	A-R	EV	Johansen	A-R	EV
100	0	2	0.5	100.0%	0.0%	0.0%	0.00	1.00	0.00
100	0	3	0.5	100.0%	0.0%	0.0%	0.00	1.00	0.00
100	0	4	0.5	100.0%	0.0%	0.0%	0.00	1.50	0.00
100	1	2	0.5	100.0%	100.0%	100.0%	1.00	1.00	1.14
100	1	3	0.5	100.0%	20.0%	40.0%	1.00	1.20	1.20
100	1	4	0.5	100.0%	0.0%	0.0%	1.00	1.63	1.00
100	2	3	0.5	100.0%	0.0%	75.0%	2.00	2.00	1.25
100	2	4	0.5	100.0%	100.0%	100.0%	2.00	2.00	1.25
100	3	4	0.5	100.0%	0.0%	83.3%	3.00	3.00	1.17
150	0	2	0.5	100.0%	0.0%	0.0%	0.00	1.00	0.00
150	0	3	0.5	100.0%	0.0%	0.0%	0.00	1.00	0.00
150	0	4	0.5	100.0%	0.0%	0.0%	0.00	1.60	0.00
150	1	2	0.5	100.0%	100.0%	100.0%	1.00	1.00	1.20
150	1	3	0.5	100.0%	25.0%	25.0%	1.00	1.25	1.00
150	1	4	0.5	100.0%	0.0%	0.0%	1.00	1.75	1.00
150	2	3	0.5	100.0%	0.0%	75.0%	2.00	2.00	1.25
150	2	4	0.5	100.0%	100.0%	100.0%	2.00	2.00	1.00
150	3	4	0.5	100.0%	0.0%	92.9%	3.00	3.00	0.93
200	0	2	0.5	100.0%	0.0%	0.0%	0.00	1.00	0.00
200	0	3	0.5	100.0%	0.0%	0.0%	0.00	1.00	0.00
200	0	4	0.5	100.0%	0.0%	0.0%	0.00	1.50	0.00
200	1	2	0.5	100.0%	100.0%	100.0%	1.00	1.00	1.00
200	1	3	0.5	100.0%	45.5%	45.5%	1.00	1.45	1.09
200	1	4	0.5	100.0%	0.0%	0.0%	1.00	1.00	1.00
200	2	3	0.5	100.0%	0.0%	66.7%	2.00	2.00	1.33
200	2	4	0.5	100.0%	83.3%	83.3%	2.00	2.17	1.17
200	3	4	0.5	100.0%	0.0%	80.0%	3.00	3.00	1.07

*Notes:* Data is generated with various values of  $T$ ,  $N$  and  $P$  for 500 simulations. The sample averages are the average values of the number of estimated cointegrations. The Accuracy is the percentage of times the true number of cointegrations is selected.

Figure 1

*Average Eigenvalues for the  $P^{\text{th}}$  and  $(P+1)^{\text{st}}$  Eigenvalue.*



*Notes:* The data is simulated using  $T = 150$ ,  $\rho = 0.0, 0.5$  one cointegrating relationship and  $N = 2, 3, 4, 5, 6, 7, 8$ . With one cointegrating relationship, the first eigenvalue will be  $\sigma_p(1)$  and the second eigenvalue is  $\sigma_p(T^{-1})$ . In all cases, both eigenvalues increase when the number of variables observed increases. When the serial correlation is low, there is a large spread between the first and second eigenvalues. As the serial correlation increases, the second eigenvalues increase rapidly for all  $N$  and the spread between the  $P$  and  $(P+1)$  eigenvalues.

## REFERENCES

- Ahn S., Horenstein A. (2010). Eigenvalue ratio test for the number of factors. *Working Paper*, Arizona State University.
- Andrews D. (1988). Laws of large numbers for dependent and non-identically distributed random variables. *Econometric Theory*. 4, 458-467.
- Bai J., Ng S. (2002). Determining the number of factors in approximate factor models. *Econometrica*. 70, 191-221.
- Binder M., Hsiao C., Pesaran M. (2005). Estimation and inference in short panel vector autoregressions with unit roots and cointegration. *Econometric Theory*. 21, 795-837.
- Bossaerts P. (1988). Common nonstationary components of asset prices. *Journal of Economic Dynamics and Control*. 12, 347-364.
- Cheung Y., Lai K. (1993). Finite-sample sizes of Johansen's likelihood ratio tests for cointegration. *Oxford Bulletin of Economics and Statistics*. 55: 313-328.
- Cragg J., Donald S. (1997). Inferring the rank of a matrix. *Journal of Econometrics*. 76, 223-250.
- Engle R., Granger C. (1987). Co-integration and error correction Representation, estimation and testing. *Econometrica* 55, 251-276
- Granger C., Weiss A. (1983). Time series analysis of error-correction models. *Studies in Econometrics, Time Series and Multivariate Statistics*. New York: Academic Press.
- Henriksen M., Isbell, J.R. (1953). On the continuity of the real roots of an algebraic equation. *Proceedings of the American Mathematical Society*. 3, 341-434.
- Ho M., Sørensen. K. (1996). Finding cointegration rank in high dimensional systems using the Johansen test: an illustration using data based Monte Carlo simulations. *The Review of Economics and Statistics*. 78, 726-732.
- Johansen S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*. 12, 231-254.
- MacKinnon J. (1996). Numerical distributions for unit root and cointegration tests. *Journal of Applied Econometrics*. 11, 601-618.



Moon H., Perron B. (2007). An empirical analysis of nonstationarity in a panel of interest rates with factors. *Journal of Applied Econometrics* 22, 383-400.

Onatski A. (2009). Testing Hypotheses about the number of factors in large factor models. *Econometrica*. 77, 1447-1479.

Phillips P. (1987). Times series regression with a unit root. *Econometrica*. 55, 277-301.

Phillips P. (1991). Optimal inference in cointegrated systems. *Econometrica*. 59, 283-306.

Phillips P., Durlauf S. (1986). Multiple time series regression with integrated processes. *The Review of Economics and Statistics*. 53, 473-495.

Phillips P., Hansen B. (1990). Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies*. 57, 473-495

Phillips P., Ouliaris S. (1990). Asymptotic properties of residual based tests for cointegration. *Economic Theory*. 58, 165-193.

Stock K., Watson M. (1989). Forecasting inflation. *Journal of Monetary Economics*. 44, 293-335

Toda H. (1994). Finite sample properties of likelihood ratio tests for cointegrating ranks when linear trends are present. *The Review of Economics and Statistics*. 76, 66-79.

APPENDIX A  
MATHEMATICAL APPENDIX

We begin by introducing some notation. First, we use  $\lambda_j(A)$  to denote the  $j^{\text{th}}$  largest eigenvalues of a symmetric matrix  $A$  and  $\mathbf{a}_j(A)$  as the eigenvector associated with the  $j^{\text{th}}$  largest eigenvalue normalized so that  $\mathbf{a}_j(A)' \mathbf{a}_j(A) = 1$ . Second, for any matrix  $X$  of full column rank, define  $P(X) = X(X'X)^{-1}X'$  and  $M(X) = I - X(X'X)^{-1}X'$ . Third, for any matrix  $X$  with  $T$  rows, the matrix  $\tilde{X}$  is used to signify a matrix where a linear time series is removed. Specifically,  $\tilde{X} \equiv M(\tau)X$ , where  $\tau = (1, 2, \dots, T)'$ .

Lemma 3.1 of Philips and Durlauf (1986) implies the following results for the data generating process in (2) or (5):

$$\begin{aligned}
\text{(i)} \quad & \frac{1}{T} \Delta Z' \tilde{F} = \frac{\Delta Z' F}{T} - \frac{1}{T} \left( \frac{\Delta Z' \tau}{T^{3/2}} \right) \left( \frac{\tau' \tau}{T^3} \right)^{-1} \left( \frac{\tau' \Delta Z}{T^{3/2}} \right) \rightarrow H_2; \\
\text{(ii)} \quad & \frac{1}{T} \tilde{F}' \tilde{F} = \frac{F' F}{T} - \frac{1}{T} \left( \frac{F' \tau}{T^{3/2}} \right) \left( \frac{\tau' \tau}{T^3} \right)^{-1} \left( \frac{\tau' F}{T^{3/2}} \right) \rightarrow H_1; \\
\text{(iii)} \quad & \frac{1}{T} \Delta Z' \tilde{U} = \frac{\Delta Z' U}{T} - \left( \frac{\Delta Z' \tau}{T^{3/2}} \right) \left( \frac{\tau' \tau}{T^3} \right)^{-1} \left( \frac{\tau' U}{T^{5/2}} \right) = O_p(1); \\
\text{(iv)} \quad & \frac{1}{T} \Delta \tilde{F}' \tilde{U} = \frac{F' U}{T} - \left( \frac{F' \tau}{T^{3/2}} \right) \left( \frac{\tau' \tau}{T^3} \right)^{-1} \left( \frac{\tau' U}{T^{5/2}} \right) = O_p(1); \\
\text{(v)} \quad & \frac{1}{T^2} \tilde{U}' \tilde{U} = \frac{U' U}{T^2} - \left( \frac{U' \tau}{T^{5/2}} \right) \left( \frac{\tau' \tau}{T^3} \right)^{-1} \left( \frac{\tau' U}{T^{5/2}} \right) = O_p(1).
\end{aligned}$$

The following lemmas are used to prove the above theorems and propositions.

Lemma 1: Suppose the two matrices  $A$  and  $B$  are symmetric of order  $p$ , then

$$\lambda_{j+k-1}(A+B) \leq \lambda_j(A) + \lambda_k(B), \quad j \leq p+1-k$$

*Proof:* See Onatski(2006) or Rao (1973, p. 68)

Lemma 2: Suppose the two matrices  $A$  and  $B$  are positive semi-definite of order  $p$ , then

$$\lambda_j(A) \leq \lambda_j(A + B), \quad j = 1, \dots, p$$

*Proof:* See Ahn and Horenstein (2011)

Lemma 3: Under Assumptions G,

$$T^{-1} Z'P(F) Z \rightarrow_p H_3$$

as  $T \rightarrow \infty$ , where  $H_3 = H_1 H_2 H_1$  and  $\text{rank}(H_3) = P$ .

*Proof:* Notice that  $T^{-1} \Delta Z'P(F) \Delta Z = (T^{-1} \Delta Z'F)^{-1} (T^{-1} F'F)^{-1} (T^{-1} F' \Delta Z)$ . Observe that

$$\frac{\Delta Z'F}{T} = \frac{1}{T} \sum_{s=1}^T B f_{s+1} f_s' - \frac{1}{T} \sum_{s=1}^T B f_s f_s' + \frac{1}{T} \sum_{s=1}^T C g_{s+1} f_s' = T^{-1} \sum_{s=1}^T [(B f_{s+1}', C g_{s+1}')' - B f_s'] f_s' \quad (\text{L3.1})$$

The matrix product is a continuous function and  $H_1$  is of full rank. Therefore, by the continuous mapping theorem we have

$$T^{-1} Z'P(F) Z \rightarrow_p H_2 H_1 H_2 = H_3 \quad (\text{L3.2})$$

Further, because  $\text{rank}(H_1) = \text{rank}(H_2) = P$  we have

$$\text{rank}(H_3) = P. \quad (\text{L3.3})$$

QED

Lemma 4: Under Assumptions A – G, with  $P \geq 1$

$$(i) \quad T^{-1} Z'P[M(F)U] Z = O_p(T^{-1})$$

$$(ii) \quad T^{-1} Z'P(U) Z = O_p(T^{-1})$$

*Proof:* The matrix product

$$T^{-1} \Delta Z'M(F)U = \frac{\Delta Z'U}{T} - \left( \frac{\Delta Z'F}{T} \right) \left( \frac{F'F}{T} \right)^{-1} \left( \frac{F'U}{T} \right) = O_p(1) \quad (\text{L4.1})$$

And

$$T^{-2}U'M(F)U = \frac{U'U}{T^2} - \frac{1}{T} \left( \frac{U'F}{T} \right) \left( \frac{F'F}{T} \right)^{-1} \left( \frac{F'U}{T} \right) = O_p(1) \quad (\text{L4.2})$$

Lemma 3.1 of Phillips and Durlauf (1986) provides the convergence results for the products involving  $U$  under Assumptions A – G provides the other convergence results; under Assumptions A – F, the matrix sums converge to functions of Brownian motion. Therefore, the matrix product is

$$T^{-1}Z'M(F)U[UM(F)U]^{-1}UM(F)Z = O_p(T^{-1}) \quad (\text{L4.3})$$

This proves part (i). The proof of part (ii) is similar. Once again, using Lemma 3.1 of Phillips and Durlauf (1986), the matrix product

$$T^{-1}\Delta Z'U = O_p(1) \quad (\text{L4.4})$$

And

$$T^{-2}UU = O_p(1) \quad (\text{L4.5})$$

The product of the matrices in (L4.4) and (L4.5) is  $O_p(1)$ . Therefore, the matrix product

$$T^{-1}\Delta Z'U(U'U)^{-1}U'\Delta Z = O_p(T^{-1}). \quad (\text{L4.6})$$

QED.

Lemma 5: Under Assumptions A – G,  $P \geq 1$ ,  $k = 1, \dots, P$  and  $l = 1, \dots, M$ , as  $T \rightarrow \infty$ ,

$$(i) \quad \lambda_k(T^{-1}\Delta Z'P(F)\Delta Z) \rightarrow_p \lambda_k(H_3) > 0;$$

$$(ii) \quad \lambda_{P+l}(T^{-1}\Delta Z'P(F)\Delta Z) \rightarrow_p 0;$$

$$(iii) \quad \lambda_l(T^{-1}\Delta Z'P[M(F)U]\Delta Z) = O_p(T^{-1}).$$

Under Assumptions A – G,  $P = 0$  and  $m = 1, \dots, N + P$ ,

$$(iv) \quad \lambda_m(T^{-1}\Delta Z'P(U)\Delta Z) = O_p(T^{-1}).$$

*Proof:* The eigenvalues of  $H_3$  are the roots of the characteristic polynomial given by the determinant of the matrix  $H_3 - I$ . It is sufficient to show that the roots of the characteristic polynomial associated with  $T^{-1} Z'P(F) Z$  converge to the roots of the characteristic polynomial associated with  $H_3$ . By Lemma 3,

$$\lim_{T \rightarrow \infty} \Pr(\|T^{-1} Z'P(F) Z - H_3\| > \epsilon) = 0, \quad \epsilon > 0. \quad (\text{L5.1})$$

Here,  $\|X\| = \max_{i,j} |X_{ij}|$  for any matrix  $X$ . Because the determinant is a continuous function of the matrix elements, the coefficients of the characteristic polynomial are also continuous functions of the matrix elements. Theorem 1 part (ii) of Henrikson and Isbell (1953) states that if a polynomial has real roots, the real roots are continuous in the coefficients. Because  $H_3$  is a positive definite matrix, all of the eigenvalues are real. Combining this result with (L5.1) proves

$$\lim_{T \rightarrow \infty} \Pr(|\lambda_j(T^{-1} \Delta Z'P(F) \Delta Z) - \lambda_j(H_3)| \geq \epsilon) = 0, \quad \forall \epsilon > 0, \quad \forall j = 1, \dots, M + P. \quad (\text{L5.2})$$

By Lemma 3,  $H_3$  is positive semidefinite with  $\text{rank}(H_3) = P$  which implies  $\lambda_k(H_3) > 0$ , for  $k = 1, 2, \dots, P$ , and  $\lambda_{P+l}(H_3) = 0$ , for  $l = 1, \dots, M$ . Combining this with (L5.2) proves parts (i) and (ii).

Next, define  $\tilde{\sim}_l$  as the eigenvector associated with the  $l^{\text{th}}$  eigenvalue of  $T^{-1} Z'P[M(F)U] Z$  normalized so that  $\tilde{\sim}_l' \tilde{\sim}_l = 1$ . The eigenvalues and eigenvectors are related through

$${}_l(T^{-1} Z'P[M(F)U] Z) = \tilde{\sim}_l'(T^{-1} Z'P[M(F)U] Z) \tilde{\sim}_l = O_p(T^{-1}). \quad (\text{L5.3})$$

The first equality is by the definition of eigenvalues and eigenvectors. The second equality comes from the normalization of the eigenvalues and Lemma 4 (i). This proves part (iii).

Define  $\tilde{\cdot}_i$  as the eigenvector associated with the 1<sup>th</sup> eigenvalue of  $T^{-1} Z'P(U) Z$  normalized so that  $\tilde{\cdot}_i' \tilde{\cdot}_i = 1$ . The eigenvalues and eigenvectors are related through

$$_i(T^{-1} Z'P(U) Z) = \tilde{\cdot}_i(T^{-1} Z'P(U) Z)_i \tilde{\cdot}_i = O_p(T^{-1}). \quad (\text{L5.4})$$

The first equality is by the definition of eigenvalues and eigenvectors. The second equality comes from the normalization of the eigenvalues and Lemma 4 (ii). This proves part (iv).

QED

Lemma 6: Under Assumptions A – G, when  $P \geq 1$ ,  $j = 1, \dots, P$  and  $k = 1, \dots, M$ ,

- (i)  $\lambda_j(\hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{ZZ}^{-1} \hat{\Omega}_{Z, \Delta Z}) = \lambda_j(\hat{\Omega}_{\Delta Z, F} \hat{\Omega}_{FF}^{-1} \hat{\Omega}_{F, \Delta Z}) + O_p(T^{-1});$
- (ii)  $\lambda_{P+k}(\hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z}) = O_p(T^{-1}).$

Under Assumptions A – F, when  $P = 0$  and for  $m = 1, \dots, M+P$ ,

- (iii)  $_m(\hat{\cdot}_{Z, Z}^{-1} \hat{\cdot}_{Z, Z}) = O_p(T^{-1})$

*Proof:* Since  $P(Z) = P([F, U]) = P(F) + P[M(F)U]$ , we have

$$\hat{\cdot}_{Z, Z}^{-1} \hat{\cdot}_{Z, Z} = Z'P(Z) Z = T^{-1} Z'P(F) Z + T^{-1} Z'P[M(F)U] Z. \quad (\text{L6.1})$$

With (L6.1), for any  $j = 1, \dots, P$ , the  $j^{\text{th}}$  largest eigenvalue of  $\hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z}$  can be written as

$$\begin{aligned} \lambda_j(\hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z}) &\leq \lambda_j(T^{-1} \Delta Z'P(F) \Delta Z) + \lambda_j(T^{-1} \Delta Z'P[M(F)U] \Delta Z) \\ &= \lambda_j(T^{-1} \Delta Z'P(F) \Delta Z) + O_p(T^{-1}) \end{aligned} \quad (\text{L6.2})$$

The first inequality is due to Lemma 1 and the second inequality is due to Lemma 5 (ii). Also, with (L6.1) and for any  $j = 1, \dots, P$ ,

$$\lambda_j(T^{-1} \Delta Z'P(F) \Delta Z) \leq \lambda_j(\hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z}) \quad (\text{L6.3})$$

This inequality is due to Lemma 2. Combining (L6.2) and (L6.3) proves Lemma 6 (i). Since  $\text{rank}(T^{-1} Z'P(F) Z) = P$ , for any  $k = 1, \dots, N$ ,

$$\lambda_{P+k}(T^{-1} \Delta Z'P(F) \Delta Z) = 0 \quad (\text{L6.4})$$

With (L6.1), for any  $k = 1, \dots, N$ , the  $(P+k)^{\text{th}}$  largest eigenvalue of  $\hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z}$  can be written as

$$\begin{aligned} & \lambda_{P+k} \left( \hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z} \right) = \lambda_{P+k} \left( T^{-1} Z'P(F) Z \right) + \lambda_{P+k} \left( T^{-1} Z'P[M(F)U] Z \right) \\ & = \lambda_{P+k} \left( T^{-1} Z'P[M(F)U] Z \right) = O_p(T^{-1}) \end{aligned} \quad (\text{L6.5})$$

The inequality is due to Lemma 1 and the equality is due to (L6.4) and Lemma 5 (ii). Further,

$$\lambda_{P+k} \left( \hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z} \right) = \lambda_{P+k} \left( T^{-1} Z'P[M(F)U] Z \right) = O_p(T^{-1}) \quad (\text{L6.6})$$

The inequality is due to Lemma 2 and the equality is due to Lemma 5 (ii). Combining (L6.5) and (L6.6) proves Lemma 6 (ii). When  $P = 0$ ,  $Z = UC'$ . Therefore, we have

$$\hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z} = \hat{\Omega}_{\Delta Z, U} \hat{\Omega}_{U, U}^{-1} \hat{\Omega}_{U, \Delta Z}.$$

Using Lemma 5 (iii) provides the result.

QED

Lemma 7: Suppose the two matrices  $A$  and  $B$  are positive definite of order  $N$ , then for  $i, j, k = 1, \dots, N$  and  $j + k \leq i + 1$ .

$$\lambda_j(AB) \leq \lambda_j(A) \lambda_k(B); \quad \lambda_{N-i+1}(AB) \geq \lambda_{N-j+1}(A) \lambda_{N-k+1}(B)$$

*Proof:* See Anderson and Dasgupta (1963)

Lemma 8: Suppose the two matrices  $A$  and  $B$  are positive definite of order  $N$ , then for  $h = 1, 2, \dots, N$ ,

$$\lambda_N(A) \lambda_h(B) \leq \lambda_h(AB) \leq \lambda_1(A) \lambda_h(B)$$



*Proof:* Set  $j = 1$ ,  $k = h$ , and  $i = h$  for Lemma 7. Then,  $\lambda_h(AB) \leq \lambda_1(A)\lambda_h(B)$ . Set  $i = N - h + 1$ ,  $j = 1$ , and  $k = N - h + 1$ . Then, we have  $\lambda_N(A)\lambda_h(B) \leq \lambda_h(AB)$ .

QED

*Proof of Theorem 1:* By Lemma 8,

$$\lambda_{P+N}(\hat{\Omega}_{\Delta Z, \Delta Z}^{-1})\lambda_r(\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z}) \leq \lambda_r(\hat{\Omega}_{\Delta Z, \Delta Z}^{-1}\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z}) \leq \lambda_1(\hat{\Omega}_{\Delta Z, \Delta Z}^{-1})\lambda_r(\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z}),$$

where  $i = 1, \dots, N + P$ . The matrix  $\hat{\Omega}_{\Delta Z, \Delta Z}^{-1}$  is positive definite and has rank  $(N+P)$ ; all of its eigenvalues are  $O_p(1)$  and positive. Thus, when  $P \geq 1$ . The convergence rates of the eigenvalues of the matrix  $\hat{\Omega}_{\Delta Z, \Delta Z}^{-1}\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z}$  are determined by the order of convergence of the eigenvalues of the matrix  $\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z}$ . Using Lemma 5 part (i) and Lemma 6 part (i),

$$\lambda_j(\hat{\Omega}_{\Delta Z, \Delta Z}^{-1}\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z}) = O_p(1), \quad j = 1, \dots, P$$

Using Lemma 8,

$$\lambda_j(\hat{\Omega}_{\Delta Z, \Delta Z}^{-1}\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z}) = O_p(T^{-1}), \quad j = P + 1, \dots, P + N$$

When  $P = 0$ ,  $Z = U$ . Therefore, we have  $\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z} = \hat{\Omega}_{\Delta Z, U}\hat{\Omega}_{U, U}^{-1}\hat{\Omega}_{U, \Delta Z}$ .

Using Lemma 6 (iii) provides the result.

QED.

*Proof of Proposition 1:* The threshold function is shown to consistently estimate the number of cointegrations. The threshold function will be larger than the  $P^{th}$  eigenvalue with probability one; alternatively, the  $P^{th}$  eigenvalue divided by the threshold function will be larger than one as  $T \rightarrow \infty$ . Because of this and Lemma 5 part (i), we have

$$\frac{\lambda_P(\hat{\Omega}_{\Delta Z, \Delta Z}^{-1}\hat{\Omega}_{\Delta Z, Z}\hat{\Omega}_{Z, Z}^{-1}\hat{\Omega}_{Z, \Delta Z})}{f(T)} = O_p(T^\alpha) \quad (\text{P1.1})$$

The ratio of the  $P^{th}$  eigenvalue and the threshold function diverges to infinity with probability one. The threshold function will be smaller than the  $(P+1)^{th}$  eigenvalue with probability one; alternatively, the threshold function divided by the  $(P+1)^{st}$  eigenvalue will be zero with probability one as  $T \rightarrow \infty$ . The smallest  $N$  eigenvalues of  $\hat{\Omega}_{\Delta Z, \Delta Z}^{-1} \hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z}$  are all positive and  $O_p(T^{-1})$ ; the threshold function is strictly positive and  $O_p(T^\alpha)$ . Because of this and Lemma 5 part (ii),

$$\frac{\lambda_{p+1} \left( \hat{\Omega}_{\Delta Z, \Delta Z}^{-1} \hat{\Omega}_{\Delta Z, Z} \hat{\Omega}_{Z, Z}^{-1} \hat{\Omega}_{Z, \Delta Z} \right)}{f(T)} = \frac{O_p(T^{-1})}{O_p(T^{-\alpha})} = O_p(T^{\alpha-1}) \quad (P1.2)$$

The ratio of the  $(P+1)^{st}$  eigenvalue and the threshold function will be smaller than one with probability one converges to zero with probability one.

QED.

*Proof of Theorem 2:* Define  $E = (e_1, e_2, \dots, e_T)'$ . From equation (11),

$$\begin{aligned} \hat{A}_p &= (Z'Z)^{-1} (Z'\Delta Z) \hat{D}_p \Xi^{1/2} \\ &= -(Z'Z)^{-1} (Z'ZA(B'A)^{-1}) \hat{D}_p \Xi^{1/2} + T^{-1} (T^{-2} Z'Z)^{-1} (T^{-1} Z'E) \hat{D}_p \Xi^{1/2} \\ &= -A(B'A)^{-1} \hat{D}_p \Xi^{1/2} + T^{-1} (T^{-2} Z'Z)^{-1} (T^{-1} Z'E) \hat{D}_p \Xi^{1/2} \\ &= -A(B'A)^{-1} \hat{D}_p \Xi^{1/2} + O_p(T^{-1}) \hat{D}_p \Xi^{1/2} \end{aligned}$$

The second equality follows when using the relationship for the differenced data in equation (7). The third equality follows because the matrices cancel each other. The last inequality follows from the limiting distribution of  $ZZ$ . The matrix  $ZZ$  can be written as  $ZZ = CUUC + CUFB + BFUC + BF FB$ .

QED.

By Lemma 3.1 of Philips and Durlauf (1986), all of the above lemma and theorems hold even if we replace  $U$  and  $F$  by  $\tilde{U}$  and  $\tilde{F}$ . Stated formally:

Lemma 9: Under Assumptions A – G,

$$T^{-1} Z'P(\tilde{F}) Z \rightarrow_p H_3 \text{ as } T \rightarrow \infty,$$

where  $\text{rank}(H_3) = P$ .

Lemma 10: Under Assumptions A – G, with  $P \geq 1$

- (i)  $T^{-1} Z'P[M(\tilde{F})\tilde{U}] Z = O_p(T^{-1});$
- (ii)  $T^{-1} Z'P(\tilde{U}) Z = O_p(T^{-1}).$

Lemma 11: Under Assumptions A – G,  $P \geq 1$ ,  $k = 1, \dots, P$  and  $l = 1, \dots, N$ , as  $T \rightarrow \infty$ ,

- (i)  $\lambda_k(T^{-1} Z'P(\tilde{F}) Z) \rightarrow_p \lambda_k(H_3) > 0;$
- (ii)  $\lambda_{P+l}(T^{-1} Z'P(\tilde{F}) Z) \rightarrow_p 0;$
- (iii)  $\lambda_l(T^{-1} Z'P[M(\tilde{F})\tilde{U}] Z) = O_p(T^{-1}).$

Under Assumptions A – G,  $P = 1$  and  $m = 1, \dots, N + P$ ,

- (iv)  $\lambda_m(T^{-1} Z'P(\tilde{U}) Z) = O_p(T^{-1}).$

Lemma 12: Under Assumptions A – G, when  $P \geq 1$ ,  $j = 1, \dots, P$  and  $k = 1, \dots, N$ ,

- (i)  $\lambda_j(\hat{\lambda}_{z, \tilde{z}}^1, \hat{\lambda}_{\tilde{z}, z}^1) = \lambda_j(\hat{\lambda}_{z, \tilde{F}}^1, \hat{\lambda}_{\tilde{F}, z}^1) + O_p(T^{-1})$
- (ii)  $\lambda_{P+k}(\hat{\lambda}_{z, \tilde{z}}^1, \hat{\lambda}_{\tilde{z}, z}^1) = O_p(T^{-1})$

Under Assumptions A – G, when  $P = 0$  and for  $m = 1, \dots, N + P$ ,

- (iii)  $\lambda_m(\hat{\lambda}_{z, \tilde{z}}^1, \hat{\lambda}_{\tilde{z}, z}^1) = O_p(T^{-1})$

*Proof of Theorem 3:* The proof follows the reasoning in the proof of Theorem 1.

*Proof of Proposition 2:* Lemma 5 part (i) states that the  $P$  largest eigenvalues of  $\hat{\lambda}_{z, \tilde{z}}^1, \hat{\lambda}_{\tilde{z}, z}^1$  will be positive while Lemma 5 part (iii) states the remaining

eigenvalues will be  $O_p(T)^{-1}$ . Using arguments similar to this in the proof of proposition 1 provide the result.

QED.