ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Determination of long-run and short-run dynamics in EC-VARMA models via canonical correlations

George Athanasopoulos, D.S. Poskitt, Farshid Vahid and Wenying Yao

November 2014

Working Paper 22/14

Determination of long-run and short-run dynamics in EC-VARMA models via canonical correlations *

George Athanasopoulos♯, D.S. Poskitt♯, Farshid Vahid♯ and Wenying Yao‡

♯ Monash University

‡ University of Tasmania

November 2, 2014

Abstract

This article studies a simple, coherent approach for identifying and estimating error correcting vector autoregressive moving average (EC-VARMA) models. Canonical correlation analysis is implemented for both determining the cointegrating rank, using a strongly consistent method, and identifying the short-run VARMA dynamics, using the scalar component methodology. Finite sample performances are evaluated via Monte-Carlo simulations and the approach is applied to model and forecast US interest rates. The results reveal that EC-VARMA models generate significantly more accurate out-of-sample forecasts than vector error correction models (VECMs), especially for short horizons.

Keywords: Cointegration, Error correction, Scalar Component Model, Multivariate Time Series.

JEL: C1, C32, C53

*Corresponding author: George Athanasopoulos, Econometrics and Business Statistics, Monash University, Caulfield East 3145, Australia, Email: george.athanasopoulos@monash.edu, Tel: +61 3 990 31075, Fax: +61 3 990 32007.

1 Introduction

Many macroeconomic time series have stochastic trends (often integrated of order 1, $I(1)$), and there are compelling theoretical reasons to expect that they are cointegrated. Cointegration refers to situations where several $I(1)$ variables share at least one *common* stochastic trend. The Granger Representation Theorem (Engle and Granger, 1987) states that all cointegrated time series have a vector error correction representation. Since most studies on cointegration are set within the context of finite-lag vector autoregressive (VAR) models, the error correcting VARs are commonly referred to as vector error correction models (VECMs). However, the Granger Representation Theorem allows for the time series of interest to have vector autoregressive moving average (VARMA) dynamics. In this paper we provide a methodology for the identification and estimation of error correcting VARMA models. While we could legitimately call such models VECMs as well, we refer to them as EC-VARMA models, and use the term "VECM" exclusively for error correcting VARs of finite order throughout the paper.

The literature on EC-VARMA models is quite limited. Lütkepohl and Claessen (1997) are the first to consider EC-VARMA models by generalising the Echelon form specification of VARMA models (Hannan and Kavalieris, 1984; Hannan and Deistler, 1988; Poskitt, 1992; Lütkepohl and Poskitt, 1996) to include an error correcting mechanism. They consider a four variable EC-VARMA model in Echelon form for US money demand and find that in general it substantially outperforms a VECM in terms of out-of-sample forecasting. They also examine a more parsimonious version of the EC-VARMA model by dropping all insignificant parameters, and this leads to an even better forecasting performance. Poskitt (2003) provides a detailed technical discussion of the theoretical results that support the EC-VARMA model in the Echelon form. Using a six variable model with U.S. macroeconomic data as an illustration he also observes a superior forecasting performance of the EC-VARMA model over a VECM but also a VARMA in levels.

Kascha and Trenkler (2014) generalize the final moving average (FMA) representation

proposed by Dufour and Pelletier (2011) to cointegrated VARMA models, and use an information criterion to choose the AR and MA orders for the cointegrated VARMA model in levels. They find promising results relative to a multivariate random walk and a standard VECM for predicting U.S. interest rates. Using the FMA representation makes their specification strategy simpler than the Echelon form, however the FMA representation is somewhat restrictive as it only focuses on a special subset of VARMA models in which the MA operator is scalar. Furthermore, the rank of the cointegration rank is taken as given. These limitations restrict the applicability of their methodology to empirical analyses.

This paper contributes to this literature in two aspects. The first is the determination of the cointegrating rank. With the exception of Poskitt (2003) the existing papers either assume this away by taking the cointegrating rank as known (Kascha and Trenkler, 2014), or use the Johansen method (Johansen, 1988) to determine the cointegrating rank (Lütkepohl and Claessen, 1997). The Johansen method is based on likelihood ratio tests for cointegration under the assumption that the data generating process (DGP) is a finite order VAR. Although this method is asymptotically justifiable for infinite order VARs as well (see Lütkepohl and Saikkonen, 1999), its finite sample performance in such situations is very poor. Here, we extend the nonparametric procedure of Poskitt (2000) (used in Poskitt 2003) to choose the cointegrating rank, to also include the case where there are no unit roots in the system. This selection procedure uses the canonical correlations between the vector time series and its first lag, and does not require any parametric assumptions about the dynamics of the underlying DGP.

The second aspect that we address is the identification and specification of the EC-VARMA model. The existing papers either rely on the identification of VARMA models in their Echelon form (Lütkepohl and Claessen, 1997), or opt for the FMA representation that is less general but easier to identify (Kascha and Trenkler, 2014). The rules for the identification of Echelon form VARMA models have often been found to be too unintuitive to understand. For example, Dufour and Pelletier (2011) state,

> "*We see that dealing with VARMA models in echelon form is not as easy as dealing with univariate ARMA models .... This might be a reason why practitioners are reluctant to employ VARMA models. Who could blame them for sticking with VAR models when they probably need to refer to a textbook to simply write down an identified VARMA representation?*"

They then suggest specifying the FMA representation that reduces the vector moving average component of a VARMA to a correlated set of scalar moving averages. These correlated univariate moving averages cannot capture the full cross covariances of a vector moving average process, and hence this method incurs some loss of information. Kascha and Trenkler (2014) base their EC-VARMA representation on the final moving average representation of Dufour and Pelletier (2011).

In this paper, we extend the Athanasopoulos and Vahid (2008*a*) method for the identification of stationary VARMA models to the partially non-stationary case. The Athanasopoulos and Vahid (2008*a*) method is based on the scalar component model (SCM) methodology, originally proposed by Tiao and Tsay (1989). One of the main contributions of this paper is that we establish the validity of the SCM methodology for non-stationary VARMA models. Specifically, we show that the SCM methodology can be applied to partially non-stationary time series in the exact same way as it is applied to stationary time series. This method identifies a dynamically complete VARMA structure that embodies at least all restrictions implied by Echelon form, which are necessary and sufficient for unique identification of a VARMA structure, and may also include some over-identifying restrictions that are supported by the data. The fact that the identification process is based on a series of moment tests makes this methodology easier to comprehend and less of a black box for applied econometricians.

In summary our paper offers a two step approach for specifying EC-VARMA models. In the first step the cointegrating rank is identified by a simple to apply strongly-consistent model selection criterion based on canonical correlations analysis. In the second step the

VARMA dynamics are identified using the SCM methodology which comprises a sequence of logical moment tests also cast in terms of canonical correlations. This makes these two steps both conceptually and procedurally consistent with each other. Once the cointegrating rank and the structure of the VARMA specification in levels are determined, the model can be formulated in its EC-VARMA form and all free parameters can be estimated using full information maximum likelihood (FIML).

We use a Monte Carlo exercise to evaluate the finite sample performance of the extended Poskitt's procedure in selecting the correct cointegrating rank and of the SCM methodology in identifying the correct VARMA structure in levels. We also examine the predictive ability of EC-VARMA models and VECMs when forecasting data generated from an EC-VARMA DGP. The computational demands of maximum likelihood estimation are impractical for Monte Carlo simulations, so we replace FIML estimation with iterative OLS (IOLS) suggested by Kapetanios (2003), which we extend to EC-VARMA models. As an empirical illustration, we use the proposed methodology to develop an EC-VARMA model for US interest rates. We find that the EC-VARMA model produces forecasts that are superior to those produced by VECMs, especially in short horizons.

The remainder of this paper is organized as follows. Section 2 defines the notation used in this paper. Sections 3 and 4 discuss the method of determining the cointegrating rank, and the SCM methodology for identifying the VARMA structure in levels for partially nonstationary time series, respectively. The Monte Carlo evaluation of these methods is conducted in Section 5. Section 6 presents an empirical application to forecasting the term structure of interest rates. Section 7 concludes.

2 Notation

The general form of a VARMA(p,q) process is

$$\boldsymbol{\Phi}_0\boldsymbol{y}_t = \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{y}_{t-p} + \boldsymbol{\Theta}_0\boldsymbol{u}_t + \boldsymbol{\Theta}_1\boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{u}_{t-q}, \tag{1}$$

where $\boldsymbol{y}_t$ is a K-dimensional time series, $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Theta}_j$ are $K \times K$ matrices of coefficients, $i = 0, 1, \ldots, p$, $j = 0, 1, \ldots, q$, and $\boldsymbol{u}_t$ is a K-dimensional vector of *i.i.d.* Gaussian white noise process with mean zero and nonsingular covariance matrix $\Sigma = \mathbb{E}(\boldsymbol{u}_t \boldsymbol{u}_t')$. The process in equation (1) can be written as

$$\boldsymbol{\Phi}(L)\, \boldsymbol{y}_t = \boldsymbol{\Theta}(L)\, \boldsymbol{u}_t,$$

where $\boldsymbol{\Phi}(L) = \boldsymbol{\Phi}_0 - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p$, $\boldsymbol{\Theta}(L) = \boldsymbol{\Theta}_0 + \boldsymbol{\Theta}_1 L + \cdots + \boldsymbol{\Theta}_q L^q$, and L is the lag operator, such that $L\, \boldsymbol{y}_t = \boldsymbol{y}_{t-1}$. The matrix polynomials satisfy

$$\det \boldsymbol{\Phi}(z) \neq 0 \quad |z| \leq 1, z \neq 1, \qquad \text{and} \qquad \det \boldsymbol{\Theta}(z) \neq 0 \quad |z| \leq 1.$$

We allow for the AR operator $\boldsymbol{\Phi}(z)$ to have roots at $z = 1$, to account for the integrated and cointegrated components of $\boldsymbol{y}_t$. Each individual time series in $\boldsymbol{y}_t$ is at most $I(1)$. The possibility that some elements in $\boldsymbol{y}_t$ may be stationary without first differencing is not excluded.

We obtain the EC-VARMA form representation from equation (1) by subtracting $\boldsymbol{\Phi}_0 \boldsymbol{y}_{t-1}$ from both sides of the equation,

$$\boldsymbol{\Phi}_0 \Delta \boldsymbol{y}_t = \boldsymbol{\Pi} \boldsymbol{y}_{t-1} + \boldsymbol{\Psi}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Psi}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{\Theta}_0 \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q}, \quad (2)$$

where $\boldsymbol{\Pi} = -(\boldsymbol{\Phi}_0 - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p)$, and $\boldsymbol{\Psi}_i = -(\boldsymbol{\Phi}_{i+1} + \cdots + \boldsymbol{\Phi}_p)$ for $i = 1, \ldots, p-1$. Denote the true cointegrating rank by ρ_0, i.e. there exist ρ_0 linear combinations of the components in $\boldsymbol{y}_t$ that are stationary, which imposes the restriction that $\operatorname{rank}(\boldsymbol{\Pi}) = \rho_0$. Hence, $\boldsymbol{\Pi}$ can be decomposed into $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both matrices of dimension $K \times \rho_0$ with full column rank. The columns of $\boldsymbol{\beta}$ represent the cointegrating vectors and $\boldsymbol{\beta}' \boldsymbol{y}_t$ represent ρ_0 long-run equilibrium relationships in $\boldsymbol{y}_t$ (Granger, 1981; Engle and Granger, 1987).

In the remainder of the paper, we assume that we have a realization of size T,

$\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$, generated from equation (1). For ease of notation, we use $\boldsymbol{y}_t$ to denote both the random vector and one realization of the random vector. $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$ and $\{\boldsymbol{y}_t\}_{t=1}^{T}$ are used interchangeably throughout the paper.

3 A non-parametric method for determining the cointegrating rank

The usual Johansen procedure for testing the cointegrating rank in the context of VECMs will be disadvantageous for EC-VARMA models due to the presence of the moving average component. Lütkepohl and Saikkonen (1999) show that the asymptotic distribution of the test statistic in Johansen's sequential likelihood ratio test remains unchanged even if the true DGP is a VARMA process, provided the order of fitted VAR is allowed to increase as the sample size increases to infinity. However, the power and size of the test is quite poor for samples of the size available for applied macroeconomic research (this was noted in Lütkepohl and Saikkonen, 1999, and is also evident from the simulation study in Section 5 below). Hence, we consider an extension of the non-parametric method of Poskitt (2000) to determine the cointegration rank.

Lütkepohl and Poskitt (1998) and Gonzalo and Pitarakis (1995) point out that the statistics for testing the cointegration rank of a multivariate system can be used to construct model selection criteria for estimating the rank consistently. The Poskitt (2000) method that we adopt here is such a model selection procedure, built on a canonical correlation based testing procedure proposed by Yang and Bewley (1996). This method does not require the true DGP to be a finite order VAR process. In fact it does not make any assumptions about the short run dynamics.

Given a sample of T observations $\{\boldsymbol{y}_t\}_{t=1}^{T}$, denote the sample squared canonical correlations between $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ (both in levels), in ascending order, as

$$\lambda_{(1),T} \leq \lambda_{(2),T} \leq \cdots \leq \lambda_{(K),T}. \tag{3}$$

For $\rho = 0, \ldots, K-1$, let $\Lambda_T(\rho)$ be the ratio of the arithmetic to the geometric mean of the $K - \rho$ largest squared canonical correlations, $\Lambda_T(\rho) = \bar{\lambda}_{\rho,T}/\bar{\lambda}^g_{\rho,T}$, where

$$\bar{\lambda}_{\rho,T} = (K-\rho)^{-1} \sum_{i=\rho+1}^{K} \lambda_{(i),T}, \qquad \text{and} \qquad \bar{\lambda}^g_{\rho,T} = \left(\prod_{i=\rho+1}^{K} \lambda_{(i),T} \right)^{1/(K-\rho)}.$$

We choose the cointegrating rank $\hat{\rho}$ to be the one that minimizes the following criterion function:

$$\zeta_T(\rho) = T(K-\rho)\ln(\Lambda_T(\rho)) + \rho(2K-\rho+1)P_T/2, \qquad \text{for} \quad \rho = 0, \ldots, K-1. \tag{4}$$

The choice of the penalty term P_T in equation (4) should satisfy the following conditions (see Poskitt, 2000, Theorem 1.2):

$$\lim_{T\to\infty} P_T/T = 0, \qquad \text{and} \qquad \lim_{T\to\infty} \ln(\ln T)/P_T = 0. \tag{5}$$

Under conditions in (5), the value of $\hat{\rho}$ that minimizes equation (4) will converge to the true cointegrating rank ρ_0 with probability 1 under certain regularity conditions. We set $P_T = \ln T$ throughout the paper.

Note that the ratio of arithmetic to geometric mean of a set of positive numbers is always greater than or equal to 1. It is equal to 1 if and only if all numbers in the set are equal to each other. Hence, the above criterion (4) is determining how many of the squared canonical correlations are equal to the largest one. However, it does not consider whether the magnitude of the largest canonical correlation is 1. Therefore, the implicit assumption of this selection criterion is that there is at least one unit root in the system. In other words, we must make sure that our vector of time series is not stationary, i.e. ρ is not equal to K. Although in practice, one always starts with univariate unit root tests, and no one would consider cointegration when all series are $I(0)$, we extend the selection procedure to ensure that the largest squared canonical correlation is equal to 1

for completeness, using the same rationale as the Poskitt (2000) method. This extension avoids K individual unit root pre-tests. The following lemma is utilized in constructing this selection criterion (see Poskitt, 2000, Lemma 1.1).

Lemma 1 *Let* $\lambda_{(i),T}$, $i = 1, \ldots, K$ *be the ordered sample squared canonical correlations in equation* (3)*, and denote their population counterparts by* $\lambda_{(1)} \leq \lambda_{(2)} \leq \cdots \leq \lambda_{(K)}$. *Then, with probability 1,*

$$\lambda_{(i),T} = \lambda_{(i)} + \boldsymbol{O}\left(\left(\frac{\ln T}{T}\right)^{1/2}\right), \qquad \textit{for} \quad i = 1, \ldots, \rho_0,$$
$$\lambda_{(i),T} = 1 + \boldsymbol{O}\left(\left(\frac{\ln \ln T}{T}\right)^{1/2}\right), \qquad \textit{for} \quad i = \rho_0 + 1, \ldots, K,$$

where $0 \leq \lambda_{(i)} < 1$, *for* $i = 1, \ldots, \rho_0$ *when the cointegrating rank in the true DGP is* ρ_0.

A significant consequence of Lemma 1 is that for large values of T, $\lambda_{(\rho_0+1),T}$, $\ldots$, $\lambda_{(K),T}$ can be arbitrarily close to unity, while $\lambda_{(1),T}, \ldots, \lambda_{(\rho_0),T}$ are strictly less than unity. Hence, to decide whether $\rho = K$, we can simply take the largest squared canonical correlation $\lambda_{(K),T}$, and compare it to $1 - \tau(\ln T/T)^{1/2}$. The decision rule is

$$\hat{\rho} = K \qquad \text{if} \qquad \lambda_{(K),T} \leq 1 - \tau\left(\frac{\ln T}{T}\right)^{1/2}, \tag{6}$$

where τ is some positive constant. We choose $\tau = 1$ for ease of exposition. According to Lemma 1, in the situation when $\rho_0 = K$, the criterion in equation (6) will choose $\hat{\rho} = \rho_0$ with probability 1. This is designed to be an extra step of the original selection criterion of Poskitt (2000). They can be used in combination, as specified in the following steps:

Step 1 For a given sample of K-dimensional time series $\{\boldsymbol{y}_t\}_{t=1}^{T}$, we first determine the sample squared canonical correlations between $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$, in ascending order, as $\lambda_{(1),T} \leq \lambda_{(2),T} \leq \cdots \leq \lambda_{(K),T}$.

Step 2 Compare $\lambda_{(K),T}$ to $1 - (\ln T/T)^{1/2}$. If $\lambda_{(K),T} \leq 1 - (\ln T/T)^{1/2}$, let $\hat{\rho} = K$. Otherwise, go to step 3.

Step 3 Construct the criterion in equation (4), and choose the cointegrating rank $\hat{\rho}$ such that

$$\hat{\rho} = arg \min_{\rho \in \{0,1,\ldots,K-1\}} \zeta_T(\rho).$$

This procedure for selecting the cointegrating rank is not confined to the class of VARMA models due to its non-parametric nature. This is preferable from both theoretical and practical perspectives, because it allows us to determine the cointegrating rank consistently without specifying the form of the short run dynamics.[1]

4 The SCM methodology

The scalar component model (SCM) was proposed by Tiao and Tsay (1989) and further developed by Athanasopoulos and Vahid (2008*a*). We demonstrate the applicability of this methodology to modelling partially stationary time series in this section. SCM VARMA representation is generally more parsimonious than the canonical Echelon form (Hannan and Kavalieris, 1984; Hannan and Deistler, 1988; Poskitt, 1992; Lütkepohl and Poskitt, 1996; Lütkepohl, 2005) since it allows for different AR and MA orders in each row. In addition, it is also based on the canonical correlation analysis, which binds the two stages of the model specification — selection of the cointegrating rank and specification of the VARMA model — nicely together.

4.1 The SCM methodology in a stationary environment

In general, the model specification of a K-dimensional zero mean VARMA(p,q) process has $(p+q)K^2$ parameters. The difficulty with VARMA modelling is that certain linear restrictions on the parameters can render a VARMA model unidentified. Take the simplest

[1] There are other system cointegration tests that use a VAR as an adjustment for short run dynamics, in order to eliminate the effect of the unknown nuisance parameters (e.g. the principal components test of Stock and Watson, 1988). Generalizing from a VAR to a VARMA adjustment may potentially improve the performances of such tests as well, but we do not explore this possibility here.

two-dimensional VARMA(1,1) process as an example,

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} \begin{pmatrix} u_{1t-1} \\ u_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}. \tag{7}$$

If $\phi_{21} = \phi_{22} = \theta_{21} = \theta_{22} = 0$, the second equation implies that $y_{2,t-1} = u_{2t-1}$ and therefore ϕ_{12} and θ_{12} cannot be uniquely identified in the first equation. In such situations, unlike finite-lag VAR models, it is not possible to fit a VARMA(1,1) model to the data and then test for statistical significance of the parameters. Moreover, it is not possible to fit VARMA models with different combinations of p and q, and then use a model selection criterion to choose among them. As a result, bespoke methodologies for VARMA modelling have been developed, one of which treats the K-dimensional VARMA model as a collection of K most parsimonious "scalar component models". A scalar component model (SCM) of orders (p_1,q_1) is a linear combination of $\boldsymbol{y}_t$ that only depends on p_1 lags of $\boldsymbol{y}_t$ and q_1 lags of $\boldsymbol{u}_t$. Hence, an SCM(p_1,q_1) can be uncovered from the implication that a linear combination of $\boldsymbol{y}_t$, $\boldsymbol{y}_{t-1}$, …, $\boldsymbol{y}_{t-p_1}$ is uncorrelated with $\boldsymbol{y}_{t-j}$ for any $j > q_1$. These SCMs are identified through a sequence of tests for moment conditions.

For a stationary VARMA process $\boldsymbol{y}_t$, let $\boldsymbol{a}'\boldsymbol{y}_t$ denote a linear combination of the components in $\boldsymbol{y}_t$. If $\boldsymbol{a}'\boldsymbol{y}_t$ is an SCM(p_1,q_1) with $p_1 < p$ and $q_1 < q$, it implies the following rank restrictions on the AR and MA coefficient matrices

$$\boldsymbol{a}' \left(\boldsymbol{\Phi}_{p_1+1}, \cdots, \boldsymbol{\Phi}_p, \boldsymbol{\Theta}_{q_1+1}, \cdots, \boldsymbol{\Theta}_q \right) = \boldsymbol{0}. \tag{8}$$

As a result of the reduced rank structure, additional restrictions are necessary in order to make $\boldsymbol{\Phi}_1$, …, $\boldsymbol{\Phi}_p$ and $\boldsymbol{\Theta}_1$, …, $\boldsymbol{\Theta}_q$ uniquely identifiable.

The SCM methodology uncovers such embedded SCMs sequentially, starting from the lowest possible orders $p_1 = 0$ and $q_1 = 0$, and continuing until K linearly independent SCMs are found. The overall orders of the VARMA model come from the highest AR and MA order among the K SCMs. Then, instead of estimating a VARMA model with

complicated cross-equation restrictions, a linear transformation of the VARMA process implied by the collection of K SCMs is estimated. Using the two-dimensional process in equation (7) as an example, if the testing procedure finds an SCM(0,0) and an SCM(1,1), then the transformed VARMA model is

$$\begin{pmatrix} 1 & 0 \\ \phi_{20} & 1 \end{pmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \theta_{11} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_{1t-1} \\ u_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}. \quad (9)$$

The left-hand side transformation matrix $\mathbf{\Phi}_0$ is not an identity matrix, but has some unknown parameters to be estimated. On the right-hand side, all of the rank restrictions implied by the SCMs are translated into zero rows in the AR and MA coefficient matrices. In addition, if $p_1 < p$ and $q_1 < q$, identifiability requires some additional zero restrictions to be placed on either the AR or the MA parameters. These restrictions are referred to as the "rule of elimination" by Tiao and Tsay (1989). In the extension of the SCM methodology in Athanasopoulos and Vahid (2008*a*), such zero restrictions are imposed on the MA parameters when needed. In the case of equation (9) we have set $\theta_{12} = 0$. All other parameters are unconstrained. The resulting uniquely identifiable form is referred to as a "canonical SCM VARMA" representation.

Once the VARMA model with the embedded SCMs is transformed into a uniquely identifiable form, and hence a canonical representation is reached, it is estimated using FIML. If the error term does not satisfy the normality assumption, the model can be estimated using generalized method of moments. The full detail of the SCM methodology is given in Athanasopoulos and Vahid (2008*a*).

4.2 The SCM methodology in a partially non-stationary environment

If $\boldsymbol{y}_t$ is a partially non-stationary VARMA process, two questions arise. Firstly, what restrictions does an SCM(p_1,q_1) imply on the EC-VARMA representation? Secondly, is it legitimate to use the same test statistic and rely on the same asymptotic distribution to uncover the embedded SCMs as in the stationary case?

Given the relationship between the coefficients of a VARMA model in levels and its corresponding error correction form, the implied restrictions of an SCM(p_1,q_1) for $\boldsymbol{y}_t$ in levels stated in (8) translate into the following restrictions on the parameters matrices

$$\boldsymbol{a}'\left(\boldsymbol{\Psi}_{p_1},\ \cdots,\ \boldsymbol{\Psi}_{p-1},\ \boldsymbol{\Theta}_{q_1+1},\ \cdots,\ \boldsymbol{\Theta}_{q}\right)=\mathbf{0}.$$

Therefore, the same linear transformation of the EC-VARMA representation produces zero rows in the AR and MA coefficient matrices or order higher than p_1-1 and q_1, respectively. Moreover, any restrictions needed to be imposed by the rule of elimination are imposed on the MA parameters reaching a canonical SCM VARMA representation. Therefore since the transformation from a VARMA model in levels to its error correction form does not alter the MA parameters, these zero restrictions will carry over to the EC-VARMA representation.

The special case is when there are one or more SCM(0,q_1) embedded in the VARMA model in levels. In this case, $\boldsymbol{a}'\boldsymbol{y}_t$ is an MA(q_1) process, and is thus stationary. As a result, $\boldsymbol{a}'\boldsymbol{y}_t$ has to be a cointegrating combination, where the vector $\boldsymbol{a}$ lies in the space spanned by the cointegrating vectors. This implies that we cannot have more than ρ_0 linearly independent SCM(0,q_1) embedded in the VARMA model.

The next question is whether we can use the canonical correlation based tests that are used for stationary VARMA processes to uncover the embedded SCMs in a partially non-stationary environment. In the stationary case, in order to make use of the moment conditions implied by the SCM(p_1,q_1) structure, we construct two $K(p_1+1)$-dimensional vectors as $\boldsymbol{Y}_{p_1,t}=\left(\boldsymbol{y}_t',\ \ \cdots,\ \ \boldsymbol{y}_{t-p_1}'\right)'$, and $\boldsymbol{Y}_{p_1,t-q_1-1}=\left(\boldsymbol{y}_{t-q_1-1}',\ \ \cdots,\ \ \boldsymbol{y}_{t-q_1-1-p_1}'\right)'$. Following previous notation, we denote the ordered population squared canonical correlation between $\boldsymbol{Y}_{p_1,t}$ and $\boldsymbol{Y}_{p_1,t-q_1-1}$ as

$$\lambda_{(1)}(p_1,q_1)\leq\lambda_{(2)}(p_1,q_1)\leq\ldots\leq\lambda_{(Kp_1+K)}(p_1,q_1).$$

These are the ordered eigenvalues of the $K(p_1+1)$-dimensional square matrix

$$\boldsymbol{A}(p_1,q_1) = \left[\mathbb{E}(\boldsymbol{Y}_{p_1,t}\boldsymbol{Y}'_{p_1,t})\right]^{-1}\mathbb{E}(\boldsymbol{Y}_{p_1,t}\boldsymbol{Y}'_{p_1,t-q_1-1}) \\ \left[\mathbb{E}(\boldsymbol{Y}_{p_1,t-q_1-1}\boldsymbol{Y}'_{p_1,t-q_1-1})\right]^{-1}\mathbb{E}(\boldsymbol{Y}_{p_1,t-q_1-1}\boldsymbol{Y}'_{p_1,t}).$$

There exist r linearly independent SCM(p_1,q_1) if and only if the multiplicity of zero eigenvalues in $\boldsymbol{A}(p_1,q_1)$ is r, i.e. $\lambda_{(1)}(p_1,q_1) = \cdots = \lambda_{(r)}(p_1,q_1) = 0$, or equivalently, $\text{rank}(\boldsymbol{A}(p_1,q_1)) = K(p_1+1) - r$.

The test for r linearly independent SCM(p_1,q_1) uses the r smallest sample squared canonical correlations $\lambda_{(1),T}(p_1,q_1)$, …, $\lambda_{(r),T}(p_1,q_1)$, which are the eigenvalues of:

$$\boldsymbol{A}_T(p_1,q_1) = (\sum_{t=1}^{T}\boldsymbol{Y}_{p_1,t}\boldsymbol{Y}'_{p_1,t})^{-1}(\sum_{t=1}^{T}\boldsymbol{Y}_{p_1,t}\boldsymbol{Y}'_{p_1,t-q_1-1}) \\ (\sum_{t=1}^{T}\boldsymbol{Y}_{p_1,t-q_1-1}\boldsymbol{Y}'_{p_1,t-q_1-1})^{-1}(\sum_{t=1}^{T}\boldsymbol{Y}_{p_1,t-q_1-1}\boldsymbol{Y}'_{p_1,t}). \quad (10)$$

The test statistic is

$$C(p_1,q_1) = -(T-p_1-q_1)\sum_{j=1}^{r}\ln\left\{1-\frac{\lambda_{(j),T}(p_1,q_1)}{d_j(p_1,q_1)}\right\}, \quad (11)$$

where

$$d_j(p_1,q_1) = 1 + 2\sum_{i=1}^{q_1}\hat{\rho}_i(\boldsymbol{\gamma}'_{(j),T}\boldsymbol{Y}_{p_1,t})\hat{\rho}_i(\boldsymbol{\delta}'_{(j),T}\boldsymbol{Y}_{p_1,t-q_1-1}),$$

$\hat{\rho}_i(\boldsymbol{x}_t)$ is the i-th lag sample autocorrelation of the process $\boldsymbol{x}_t$, and $\boldsymbol{\gamma}_{(j),T}$ and $\boldsymbol{\delta}_{(j),T}$ are the canonical covariates corresponding to the eigenvalue $\lambda_{(j),T}(p_1,q_1)$. Asymptotically, the test statistic follows a χ^2 distribution with r^2 degrees of freedom under the null hypothesis of r zero eigenvalues.

In the partially non-stationary case, we consider a different normalization of the process $\boldsymbol{y}_t$. For a time series $\boldsymbol{y}_t$ that comes from a VARMA(p,q) process, there exists a $K \times K$ nonsingular transformation matrix $\boldsymbol{H}$ such that $\boldsymbol{H}\boldsymbol{y}_t = (\boldsymbol{n}'_t,\ \boldsymbol{s}'_t)'$, where $\boldsymbol{s}_t$ is a purely stationary process (see Poskitt, 2000, for an example of such a transformation). Given

the cointegrating relationship in $\boldsymbol{y}_t$, it follows that $\boldsymbol{s}_t$ has dimension $\rho_0 \times 1$. Thus, the difference stationary component $\boldsymbol{n}_t$ has dimension $(K - \rho_0) \times 1$. Let

$$\boldsymbol{x}_t = \boldsymbol{G}_T \boldsymbol{H} \boldsymbol{y}_t = \begin{pmatrix} \boldsymbol{n}_t / T^{1/2} \\ \boldsymbol{s}_t \end{pmatrix}, \qquad \text{where} \quad \boldsymbol{G}_T = \begin{pmatrix} \boldsymbol{I}_{K-\rho_0}/T^{1/2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\rho_0} \end{pmatrix}$$

and $\boldsymbol{I}_m$ the $m \times m$ identity matrix. We construct $K(p_1+1)$-dimensional vectors $\boldsymbol{X}_{p_1,t}$ and $\boldsymbol{X}_{p_1,t-q_1-1}$ as follows

$$\boldsymbol{X}_{p_1,t} = \begin{pmatrix} \boldsymbol{n}_t / T^{1/2} \\ \tilde{\boldsymbol{s}}_{p_1,t} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}_T \boldsymbol{H} \boldsymbol{y}_t \\ \boldsymbol{y}_t - \boldsymbol{y}_{t-1} \\ \vdots \\ \boldsymbol{y}_{t-(p_1-1)} - \boldsymbol{y}_{t-p_1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}_T \boldsymbol{H} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I}_K & -\boldsymbol{I}_K & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I}_K & -\boldsymbol{I}_K \end{pmatrix} \boldsymbol{Y}_{p_1,t} = \boldsymbol{F} \boldsymbol{Y}_{p_1,t}, \tag{12}$$

and $\boldsymbol{X}_{p_1,t-q_1-1} = \left(\boldsymbol{n}'_{t-q_1-1}/T^{1/2},\ \tilde{\boldsymbol{s}}'_{p_1,t-q_1-1}\right)' = \boldsymbol{F}\, \boldsymbol{Y}_{p_1,t-q_1-1}$. Note that $\boldsymbol{F}$ is nonsingular. It follows from equation (12) that $\tilde{\boldsymbol{s}}_{p_1,t}$ and $\tilde{\boldsymbol{s}}_{p_1,t-q_1-1}$ — the last $(Kp_1 + \rho_0)$ components of $\boldsymbol{X}_{p_1,t}$ and $\boldsymbol{X}_{p_1,t-q_1-1}$ — are strictly stationary.

The sample squared canonical correlations between $\boldsymbol{X}_{p_1,t}$ and $\boldsymbol{X}_{p_1,t-q_1-1}$ are the eigenvalues of

$$\begin{aligned} \boldsymbol{A}_T^x(p_1,q_1) = (\sum_{t=1}^{T} \boldsymbol{X}_{p_1,t}\boldsymbol{X}'_{p_1,t})^{-1}(\sum_{t=1}^{T} \boldsymbol{X}_{p_1,t}\boldsymbol{X}'_{p_1,t-q_1-1}) \\ (\sum_{t=1}^{T} \boldsymbol{X}_{p_1,t-q_1-1}\boldsymbol{X}'_{p_1,t-q_1-1})^{-1}(\sum_{t=1}^{T} \boldsymbol{X}_{p_1,t-q_1-1}\boldsymbol{X}'_{p_1,t}). \end{aligned} \tag{13}$$

We denote these eigenvalues ordered from smallest to largest by

$$\lambda^x_{(1),T}(p_1,q_1) \leq \lambda^x_{(2),T}(p_1,q_1) \leq \ldots \leq \lambda^x_{(Kp_1+K),T}(p_1,q_1),$$

and their corresponding eigenvectors by $\boldsymbol{\zeta}_{(j),T}, j = 1, \ldots, Kp_1 + K$. The following lemma states a set of results that lead us to the conclusion that the same testing procedure can

be used for identifying SCMs in the partially non-stationary case.

Lemma 2 *Let* $\{\lambda_{(j),T}, \boldsymbol{\gamma}_{(j),T}, j = 1, \ldots, Kp_1 + K\}$ *and* $\left\{\lambda^x_{(j),T}, \boldsymbol{\zeta}_{(j),T}, j = 1, \ldots, Kp_1 + K\right\}$ *be eigenvalue, eigenvector pairs of matrices* $\boldsymbol{A}_T(p_1, q_1)$ *and* $\boldsymbol{A}^x_T(p_1, q_1)$ *defined in equations* (10) *and* (13)*. Then,*

1. $\lambda^x_{(j),T}(p_1, q_1) = \lambda_{(j),T}(p_1, q_1)$ *for* $j = 1, \ldots, Kp_1 + K$ *and all* T*;*

2. $\boldsymbol{\zeta}'_{(j),T}\boldsymbol{X}_{p_1,t} = \boldsymbol{\gamma}'_{(j),T}\boldsymbol{Y}_{p_1,t}$ *for* $j = 1, \ldots, Kp_1 + K$ *and all* T*;*

3. *When* $\boldsymbol{Y}_{p_1,t}$ *and* $\boldsymbol{Y}_{p_1,t-q_1-1}$ *do not overlap, the largest* $K - \rho_0$ *eigenvalues converge to 1, and the rest converge to probability limits that are less than 1, and the rate of convergence of the largest* $K - \rho_0$ *eigenvalues to 1 is faster than that of the rest.*

4. *When* $\boldsymbol{Y}_{p_1,t}$ *and* $\boldsymbol{Y}_{p_1,t-q_1-1}$ *do have* m *common elements,* m *eigenvalues will be identically equal to 1. Among the remaining eigenvalues,* $K - \rho_0$ *converge to 1 and the rest converge to probability limits that are less than 1, and the rate of convergence of the largest* $K - \rho_0$ *eigenvalues to 1 is faster than that of the rest.*

5. *Every eigenvalue that has probability limit less than 1, converges to a population squared canonical correlations between* $\tilde{\boldsymbol{s}}_{p_1,t}$ *and* $\tilde{\boldsymbol{s}}_{p_1,t-q_1-1}$*. Moreover, the corresponding canonical covariate* $\boldsymbol{\gamma}'_{(j),T}\boldsymbol{Y}_{p_1,t}$ *in the limit will only involve* $\tilde{\boldsymbol{s}}_{p_1,t}$*, the stationary components of* $\boldsymbol{Y}_{p_1,t}$*. Then* $\boldsymbol{\delta}'_{(j),T}\boldsymbol{Y}_{p_1,t-q_1-1}$*, the corresponding combination of* $\boldsymbol{Y}_{p_1,t-q_1-1}$ *that has the highest correlation with* $\boldsymbol{\gamma}'_{(j),T}\boldsymbol{Y}_{p_1,t}$ *will also only involve* $\tilde{\boldsymbol{s}}_{p_1,t-q_1-1}$ *in the limit.*

Proof. 1 and 2 are immediate consequences of

$$\boldsymbol{A}^x_T(p_1, q_1) = (\boldsymbol{F}')^{-1}\boldsymbol{A}_T(p_1, q_1)\boldsymbol{F}'.$$

3, 4 and 5 are straightforward generalisations of Lemma 1. ■

The test for r linearly independent SCM(p_1,q_1) is implemented as a test for the joint significance of the r smallest eigenvalues in $\boldsymbol{A}_T(p_1, q_1)$. Lemma 2 demonstrates that under

the null, these r eigenvalues are solely determined by the stationary components of $\boldsymbol{Y}_{p_1,t}$. Thus, the test would be the same as testing for r zero canonical correlations between $\tilde{\boldsymbol{s}}_{p_1,t}$ and $\tilde{\boldsymbol{s}}_{p_1,t-q_1-1}$. Moreover, $\boldsymbol{\gamma}'_{(j),T}\boldsymbol{Y}_{p_1,t}$ and $\boldsymbol{\delta}'_{(j),T}\boldsymbol{Y}_{p_1,t-q_1-l}$ for $j = 1,\ldots,r$ are consistent estimators of covariates corresponding to the r smallest canonical correlations between $\tilde{\boldsymbol{s}}_{p_1,t}$ and $\tilde{\boldsymbol{s}}_{p_1,t-q_1-1}$. Therefore, the test statistic $C(p_1,q_1)$ defined in (11) will have the same asymptotic distribution as in the stationary case.

5 Monte Carlo simulation

We use a Monte Carlo simulation to assess the performance of the methods for selecting the cointegrating rank, and for identifying the SCM structure presented in Sections 3 and 4. We also evaluate the predictive ability of the identified EC-VARMA models versus alternative VECM specifications for which the cointegrating rank is selected by the Johansen procedure. The lag length for VAR in levels is selected by the AIC, HQ or BIC as is usual in practice. The DGP we use in the simulation is a 3-dimensional VARMA(1,1) in levels,

$$\boldsymbol{y}_t = \begin{pmatrix} 0.75 & 0.25 & 0 \\ 0.11 & 0.89 & 0 \\ -0.1 & 0.1 & 1 \end{pmatrix} \boldsymbol{y}_{t-1} + \boldsymbol{u}_t + \begin{pmatrix} -0.35 & 0.2 & -0.54 \\ 0.7 & 0.5 & 0.1 \\ -0.4 & 0.75 & 0.6 \end{pmatrix} \boldsymbol{u}_{t-1}, \tag{14}$$

where $\boldsymbol{u}_t$ is *i.i.d.* $\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_3)$. The AR and MA orders are both one to simplify the illustration. The EC-VARMA(0,1) representation is $\Delta\boldsymbol{y}_t = \boldsymbol{\Pi}\boldsymbol{y}_{t-1} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1\boldsymbol{u}_{t-1}$, where $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, $\boldsymbol{\alpha} = (-0.25, 0.11, -0.1)'$ and $\boldsymbol{\beta} = (1, -1, 0)'$. Hence the true cointegrating rank is $\rho_0 = 1$. All three eigenvalues of $\boldsymbol{\Theta}_1$ are close to 0.8, indicating the presence of a relatively strong propagation mechanism in the MA dynamics.

5.1 Selection of cointegrating rank, lag length and the SCM structure

We determine the cointegrating rank using the extended Poskitt (2000) procedure developed in Section 3. We simulate 1000 replications for four different sample sizes: $T = 100$, 200, 400 and 1000. As shown in Figure 1, the proposed procedure correctly selects the true cointegrating rank, $\hat{\rho} = \rho_0 = 1$, 100% of the time, even for samples with 100 observations.[2] The non-parametric nature of the Poskitt (2000) procedure and its extension proposed here allows for this procedure to be used independently of the assumed underlying model. In contrast, the Johansen procedure is dependent on the lag length of the assumed underlying cointegrated VAR. Figure 2 plots the distribution of the estimated lag lengths for VARs in levels for different sample sizes. The maximum lag length is set to 20. All three of the information criteria choose longer lags as the sample size T increases. The AIC has the tendency to choose very long lags for $T = 100$.

Insert Figure 1 here.

Insert Figure 2 here.

Conditional on the selected lag lengths, the distributions of the estimated cointegrating rank $\hat{\rho}$ selected by the Johansen procedure are also plotted in Figure 1. It is evident that the actual size of the Johansen procedure is far from its nominal size of 5% even for $T = 1000$. This result is in accord with Lütkepohl and Saikkonen (1999) when the DGP is a cointegrated VARMA process. Furthermore, for $T = 100$, the Johansen procedure coupled with AIC chooses the correct specification of the cointegrating rank for below 60% of the time. This reveals that the cointegrating rank estimated using the Johansen procedure depends crucially on the lag length of the VECM. Clearly the evidence in this

[2] We have also performed more comprehensive simulations for selecting the cointegrating rank with the Poskitt procedure using 100 different DGPs and 1000 simulations for each DGP, and we find that the procedure can choose the true cointegrating rank for at least 95% of the time even when $T = 100$. These simulation results are available upon request.

simulation supports using the modified Poskitt (2000) procedure instead of the Johansen method for correctly selecting the cointegrating rank.

We next evaluate the performance of the SCM methodology to identify the VARMA dynamics. Athanasopoulos and Vahid (2008*a*) and Athanasopoulos et al. (2012) have found that the SCM identification procedure is quite successful in identifying VARMA structures in a stationary setting. We find similar results here for data generated from a cointegrated VARMA process. Due to the manual implementation of the SCM identification process for each simulated sample for this part of the simulation we restrict the number of iterations to 100 for each of the sample sizes: T = 100, 200, 400 and 1000. We present the results in Table 1. The SCM structure implied by the DGP of equation (14) is three SCM(1,1). The testing procedure is able to identify the correct structure for more than 95% of the time, even for $T = 100$.

5.2 Estimation of the EC-VARMA models

The common approach in the literature is to estimate all of the unknown parameters in the EC-VARMA model simultaneously using FIML. However, it is infeasible and computationally inefficient to use FIML in Monte Carlo simulations. Therefore, we use an iterative OLS (IOLS) procedure to estimate the EC-VARMA models in the simulation. This procedure extends the IOLS method suggested by Kapetanios (2003) for stationary VARMA models.

Given an identified EC-VARMA($p-1$,q) model with K embedded SCMs, we only include variables with non-zero coefficients as the explanatory variables on the right-hand side. In each iteration, we replace the lagged error terms $\boldsymbol{u}_{t-1}, \ldots, \boldsymbol{u}_{t-q}$ by the lagged residual series obtained from the last iteration. The initial estimate of the error sequence is obtained from the residual of a finite VAR fitted to the simulated data. We let the lag length of the VAR be $\lceil \ln T \rceil$ as was suggested by Lütkepohl and Poskitt (1996). Denote the estimated cointegrating rank as $\hat{\rho}$, the cointegrating vectors are re-estimated in each iteration. We calculate the partial canonical correlations between $\Delta \boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ after

controlling for $\Delta \boldsymbol{y}_{t-1}$, ..., $\Delta \boldsymbol{y}_{t-p+1}$, and $\hat{\boldsymbol{u}}_{t-1}$, ..., $\hat{\boldsymbol{u}}_{t-q}$. The canonical covariates corresponding to the largest $\hat{\rho}$ squared partial canonical correlations are taken as the estimated cointegrating vectors. The rest of the unknown parameters are then estimated using OLS regression with the error correction terms and the lagged error terms taken as known.

If the iterative procedure converges such that the estimated covariance matrix of the residual sequence stabilizes, then the OLS estimates of the coefficients from the last iteration are adopted. However, if the convergence condition cannot be achieved within a pre-specified maximum number of iterations, we suggest to use the estimator given by Hannan and Kavalieris (1984) to produce the final estimates. Appendix A presents detailed steps for the implementation of the IOLS procedure.

5.3 A forecast evaluation of EC-VARMA models versus VECMs

For each sample path, we generate forecasts from the identified EC-VARMA specifications and the VECMs. We select the cointegrating rank for the VECMs using the Johansen procedure (denoted as $\hat{\rho}_J$ in Tables 2 and 3), as well as the modified Poskitt (2000) approach (denoted as $\hat{\rho}_P$). The modified Poskitt (2000) approach is used by default for selecting the cointegrating rank of the EC-VARMAs.

Insert Table 2 here.

Insert Table 3 here.

We compare the forecasts generated by the estimated models to the theoretical forecasts, i.e., the forecasts generated from the EC-VARMA(0,1) DGP with the true parameters. We refer to this as the "oracle". We evaluate the forecast performance over horizons $h = 1, \ldots, 24$, and consider two measures of forecasting accuracy — the trace of mean squared forecast errors (tr(MSFE)) for $\boldsymbol{y}_t$ in levels, and the generalized forecast error second moment (GFESM). The GFESM is the determinant of the forecast error second

moment matrix pooled across all horizons,

$$\text{GFESM}_h = \left(\det\left(\mathbb{E}[\text{vec}(\mathbf{e}_1,\ldots,\mathbf{e}_h)\ \text{vec}(\mathbf{e}_1,\ldots,\mathbf{e}_h)']\right)\right)^{1/h},$$

where $\mathbf{e}_i$ is the $K\times 1$ dimensional vector of the i-th step ahead forecast error, $i=1,\ldots,h$. The main advantage of GFESM is that it is invariant to non-singular, scale preserving linear transformations for all forecast horizons (see Clements and Hendry, 1993, for details).

Tables 2 and 3 present the percentage differences in tr(MSFE) and GFESM between the estimated models and the "oracle". For example, in the first row of Table 2, the entry in the column under "EC-VARMA" denotes that for $h=1$ the tr(MSFE) from the EC-VARMA models estimated using IOLS, is 16.3% larger than the "oracle" for samples of $T=100$. Similar interpretations can be drawn from Table 3 for the GFESM. Entries in bold indicate the model specification that is most accurate amongst the estimated models for that forecast horizon. A negative entry shows an improvement over the "oracle".

The results in Tables 2 and 3 show that in general, given the typical sample sizes available for macroeconomic data, approximating an EC-VARMA process with a VECM generates a considerable loss in forecast accuracy. For relatively small sample sizes, the losses are quite substantial for short forecast horizons.

The columns labeled $\hat{\rho}_J$ and $\hat{\rho}_P$ in Tables 2 and 3 allow us to examine the effects on the forecasting accuracy of using different cointegrating ranks. The lag lengths of VECMs are selected by the same information criteria, but the cointegrating ranks are chosen by the Johansen procedure and the extended Poskitt (2000) method, respectively. These tables show that the use of the extended Poskitt (2000) method produces smaller forecast error, especially when the sample size is small.

6 Term structure of interest rates

It is commonly accepted that interest rates with different maturities are cointegrated (see Hall et al., 1992). The cointegrating vector between any two interest rate series should be

close to (1, -1), i.e., the interest rate spreads should be stationary, despite the fact that most interest rates are regarded as $I(1)$ series. Many studies of interest rates have been conducted within the VECM framework. To name one among others, Hall et al. (1992) find that yields to maturity of US treasury bills specify an error correction model with post-1982 data, which proves to be useful in forecasting changes in yields.

In a recent paper Kascha and Trenkler (2014) show that a cointegrated VARMA model generates superior forecasts for US interest rates. Our application differs from theirs in several aspects. First, Kascha and Trenkler (2014) extend the FMA representation of Dufour and Pelletier (2011) to specify their VARMA model. This approach is simpler but less general and less parsimonious than the SCM representation employed here. Moreover, Kascha and Trenkler (2014) only take the cointegrating rank as given ($\rho = K - 1$) for their forecast evaluation, whereas we test for cointegration rank for each sample.

6.1 Data

We use monthly data of the US federal funds rate, and 3-month and 6-month treasury bill rates to form a three-variable system. Let $\boldsymbol{y}_t = (ff_t, i3_t, i6_t)'$. The available sample period is from 1958:12 to 2011:09, leads to a total of 634 observations. Figure 3 plots the three interest rate series over the entire sample period. The movements in the three series clearly share a similar pattern, especially for the 3-month and 6-month treasury bill rates.[3]

Insert Figure 3 here.

We use the first 400 observations as the initial estimation sample to forecast future interest rates for up to 12-step ahead. We use an expanding window which adds one observation to the estimation sample at a time, and repeat the same forecasting exercise 222

[3]One may want to drop the observations during the last global financial crisis (2008:01-2011:09, the last 45 observations), due to the possibility of a structural break. We experimented with this shorter sample as well, and it produces qualitatively similar results. The forecast errors are smaller in almost all cases, but the ranking among the competing models does not change.

times until the end of the sample.[4] We re-identify and re-estimate the model specification for each estimation sample. We calculate the tr(MSFE), the determinant of the MSFE (det(MSFE)) and the GFESM over the 222 estimation windows.

6.2 Selection of cointegrating rank

Theoretically for a K-dimensional system of interest rates the cointegrating rank should be $\rho = K - 1$. In our system the theoretically supported rank is $\rho = 2$. As shown by the last row in Table 4, the modified Poskitt (2000) procedure chooses $\hat{\rho} = 2$ consistently for all 222 estimation samples. Table 4 also shows the rank chosen by the Johansen procedure conditional on the lag length selected for a VAR in levels by the three model selection criteria. The maximum lag length set to 24, i.e., two years for monthly data. The results show that the AIC chooses a VAR(21) in levels 100% of the time. Allowing for this very long lag length leads the Johansen procedure to select a cointegrating rank $\hat{\rho} = 1$, 52% of the time and $\hat{\rho} = 0$ the rest of the time. The Johansen procedure coupled with the AIC never selects the theoretically supported rank of $\hat{\rho} = 2$. In contrast, the shorter lag lengths selected by the BIC and the HQ lead to selecting $\hat{\rho} = 2$ almost 100% of the time.

Insert Table 4 here.

6.3 Canonical SCM VARMA identification

We use the SCM methodology to determine the orders and the corresponding canonical structure of the VARMA model in levels. Conditional on the identified tentative overall VARMA(1,1) structure for $\boldsymbol{y}_t$, we search for each individual SCM. Starting from the most parsimonious SCM(0,0), the underlying SCMs identified are SCM(1,1), SCM(1,1) and SCM(1,0). After testing the SCM structure of the sub-systems and imposing identification restrictions on $\boldsymbol{\Phi}_0$ (see Athanasopoulos and Vahid, 2008*a*), the canonical SCM VARMA

[4] We have also implemented a rolling window to generate forecasts, which can account for the possible structural breaks over the time span that we consider. The results are very similar to what we report here using an expanding window, and hence are omitted here. Those results are available upon request.

model has the following error correction representation,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a_0 & 0 & 1 \end{pmatrix} \Delta \boldsymbol{y}_t = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \\ \alpha_{31} & \alpha_{32} \end{pmatrix} \begin{pmatrix} 1 & \beta_{12} \\ \beta_{21} & 1 \\ \beta_{31} & \beta_{32} \end{pmatrix}' \boldsymbol{y}_{t-1} + \boldsymbol{u}_t + \begin{pmatrix} \theta_{11}^1 & \theta_{12}^1 & \theta_{13}^1 \\ \theta_{21}^1 & \theta_{22}^1 & \theta_{23}^1 \\ 0 & 0 & 0 \end{pmatrix} \boldsymbol{u}_{t-1}. \quad (15)$$

We find exactly the same SCM structure for all 222 estimation windows. There are no zero restrictions imposed on $\boldsymbol{\Phi}_1$ by the canonical SCM structure. Hence, the rank restrictions on $\boldsymbol{\Pi}$ only come from the cointegrating relationships. We estimate equation (15) using FIML. In order to provide good initial estimates for maximum likelihood estimation, we put the algorithm of Hannan and Rissanen (1982) into the context of EC-VARMA models, and use their estimates as the starting values for the maximum likelihood iteration.

6.4 Forecast evaluation of the interest rates

The forecast evaluation results are presented in Table 5. We set the identified SCM EC-VARMA specification (15) with the theoretically supported cointegrating vectors as our benchmark. The cointegrating vectors are specified as

$$\boldsymbol{\beta} = \left[\begin{pmatrix} 1, & -1, & 0 \end{pmatrix}', \quad \begin{pmatrix} 1, & 0, & -1 \end{pmatrix}' \right]. \quad (16)$$

We evaluate the forecasting performance of the identified EC-VARMA model (15) when the cointegrating rank is specified by the modified Poskitt method, and the cointegrating vectors are estimated from the data. We denote these as EC-VARMA$_P$ (last column of Table 5). It is worth pointing out that the only difference between EC-VARMA$_P$ and the benchmark is solely due to the estimation of the two cointegrating vectors as the modified Poskitt method selects $\hat{\rho} = 2$, 100% of the time.

For the alternative VECM specifications, we impose the theoretically supported cointegrating vectors (16) as shown in columns 2-4 in Table 5. We also estimate the cointegrating relationships after the cointegrating rank has been selected by the Johansen procedure,

which is denoted as VECM$_J$ in columns 5-7 in Table 5.

The entries in Table 5 show the percentages by which the measures of forecasting accuracy for each model specification are larger than the benchmark EC-VARMA with theoretical cointegration imposed. Hence a negative entry indicates an improvement in forecasting accuracy over the benchmark. Entries in bold indicate the largest percentage gain over the benchmark for that forecast horizon. If no entries in bold exist, the benchmark generates the most accurate forecasts for the horizon.

The results clearly show that for all three measures of forecasting accuracy, EC-VARMA models are the more accurate than VECMs. EC-VARMA models are substantially more accurate than VECMs especially for the shorter forecast horizons. Considering the tr(MSFE), the EC-VARMAs having imposed the theoretically supported cointegrating vectors are the most accurate whereas considering the det(MSFE) and the GFESM the ECVARMAs having estimated the cointegrating relationships are the most accurate.

In evaluating the forecasting performance of the VECM specifications, the results clearly show that selecting lag length using the AIC generates the least accurate forecasts. It is most likely that the large estimation error associated with the large number of parameters in these specifications causes this result. Selecting the lag length using the BIC generates the most accurate forecasts for the VECMs. For all the VECMs, it seems that imposing the theoretically supported cointegrating vectors improves, or at the very least does not hinder their forecasting accuracy especially for short forecast horizons.

Insert Table 5 here.

In columns 5-7 of Table 5, we also report Diebold-Mariano (DM) test results (Diebold and Mariano, 1995; West, 1996; Giacomini and White, 2006) for comparing the predictive accuracy between the EC-VARMA$_P$ models and the alternative VECM$_J$ specifications, both of which use estimated cointegratiion vectors. These specifications are the ones most commonly implemented in practice. We test for the equality of MSFE of each pair of

forecasts for each series ff_t, $i6_t$ and $i3_t$ individually, and also for the tr(MSFE) using a generalisation of the DM test for multivariate models. The test hypotheses for the individual series are

$$H_0 : \mathbb{E}\left[\boldsymbol{e}_{1,i,h}^2\right] - \mathbb{E}\left[\boldsymbol{e}_{2,i,h}^2\right] = 0, \quad \text{against} \quad H_1 : \mathbb{E}\left[\boldsymbol{e}_{1,i,h}^2\right] - \mathbb{E}\left[\boldsymbol{e}_{2,i,h}^2\right] < 0,$$

and for the tr(MSFE),

$$H_0 : \mathbb{E}\left[\sum_{i=1}^{3} \boldsymbol{e}_{1,i,h}^2\right] - \mathbb{E}\left[\sum_{i=1}^{3} \boldsymbol{e}_{2,i,h}^2\right] = 0, \quad \text{against} \quad H_1 : \mathbb{E}\left[\sum_{i=1}^{3} \boldsymbol{e}_{1,i,h}^2\right] - \mathbb{E}\left[\sum_{i=1}^{3} \boldsymbol{e}_{2,i,h}^2\right] < 0,$$

where $\boldsymbol{e}_{1,i,h}$ and $\boldsymbol{e}_{2,i,h}$ denote the h-step ahead forecast errors of the i-th component of $\boldsymbol{y}_t = (ff_t, i6_t, i3_t)'$ for $i = 1, 2, 3$, generated from the estimated EC-VARMA$_P$ and VECM$_J$ specifications respectively. Newey-West standard errors are used in these tests.

Table 5 presents the results for the tr(MSFE) using a superscript symbol for each tr(MSFE) entry and the individual test results using three comma-separated symbols (one for each series) below each tr(MSFE) entry. A ** (*, †) indicates the rejection of the null hypothesis in favor of the one sided alternative at 1% (5%,10%) level of significance.

The tr(MSFE) generated from the estimated EC-VARMA$_P$ models is significantly lower than the tr(MSFE) of VECM$_J$ specifications for short horizons. Specifically, it is significantly lower than: VECM$_J$ with BIC up to 3-step ahead at 1% significance level; VECM$_J$ with HQ up to 4-step ahead at 10% significance level; and VECM$_J$ with AIC up to 6-step ahead at 5% significance level. Examining the testing results for individual series, these statistically significant differences seem to be driven by more accurate forecasts generated from the EC-VARMA$_P$ for ff_t, although statistically significant differences can also be seen for $i6_t$ and the $i3_t$. It is important to highlight here that there are some cases for the individual series where the EC-VARMA$_P$ specifications generated a larger MSFE than the VECMs. However, in each of these cases, the differences are all statistically insignificant.

7 Conclusion

Applied researchers have tended to favor finite order VARs and VECMs rather than the more general and flexible VARMA models in macroeconomic modelling. This is due mainly to the specification and estimation difficulties encountered with VARMA models. In this paper we combine cointegration relationships among non-stationary time series with VARMA models and address the specification of EC-VARMA model from two aspects. We first extend the non-parametric approach of Poskitt (2000) for determining the cointegrating rank to include the possibility of $I(0)$ variables. Our Monte Carlo simulations show that this extended version has good finite sample performance in determining the true cointegrating rank. Moreover it has a significant advantage over the traditional Johansen procedure when the underlying data generating process is a cointegrated VARMA.

Secondly, we demonstrate that the testing procedure of the SCM methodology for identifying the canonical SCM structure of stationary VARMA models (Athanasopoulos and Vahid, 2008*a*) is valid in the partially non-stationary environment. Our Monte-Carlo results show that the SCM methodology performs well in identifying VARMA structures in a partially non-stationary setting, similarly to the results of Athanasopoulos and Vahid (2008*a*) and Athanasopoulos et al. (2012) for a stationary setting.

Athanasopoulos and Vahid (2008*b*) provide extensive evidence that in a stationary environment not accounting for moving average dynamics hinders significantly the forecasting accuracy of multivariate models. In this paper we ask a similar question in a partially non-stationary setting using Monte Carlo simulations. The answer is again the same. Approximating EC-VARMA dynamics with long order VECMs significantly hinders forecasting accuracy especially for short horizons. Using our extension to the Poskitt (2000) method to determine the cointegrating rank has significant impact in improving forecasting accuracy, especially in small samples.

In our empirical application we model and forecast the term structure of US interest rates. The results provide further evidence supporting the forecast superiority of EC-

VARMA models over VECMs. We estimate EC-VARMA models and VECMs with the cointegrating relationships either estimated from the data or dictated by economic theory. The EC-VARMA models with either estimated or theoretical cointegrating relationships always produce more accurate forecasts than VECMs. Diebold-Mariano tests show that these differences are statistically significant, especially in the short run.

This paper contributes to the growing body of literature on the identification and estimation of VARMA models. It suggests that EC-VARMA models can be both beneficial and relatively straightforward to estimate using the methods we explore here. The empirical evidence in favor of both stationary and partially non-stationary VARMA models over VARs is mounting. We hope that this evidence coupled with the methodological improvements in overcoming modelling complexities will lead to VARMA models being utilized more comprehensively by the applied econometrician in macroeconomic modelling and forecasting.

Appendix A: The iterative OLS procedure

The VARMA(p,q) model to be estimated is in the error correction form:

$$\boldsymbol{\Phi}_0 \Delta \boldsymbol{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{y}_{t-1} + \boldsymbol{\Psi}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Psi}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q}, \quad (17)$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both $K \times \hat{\rho}$ dimensional matrices of full column rank, and there are proper restrictions imposed on the coefficient matrices in order to ensure unique identification. The traditional approach in the literature is to estimate all unknown parameters simultaneously using FIML, which is the exact method that we use for the empirical application in Section 6. However, it is both infeasible and computationally inefficient to use FIML with large scale simulations in Section 5, because it may occasionally fail to converge. Thus, we use the following iterative procedure to estimate model (17) in the Monte Carlo simulation. This procedure is built upon the iterative OLS (IOLS) estimation suggested by Kapetanios (2003) for stationary VARMA models.

The initial estimate of the error sequence $\hat{\boldsymbol{u}}_t^0$ is obtained from the residual of a VAR, where the lag length of the VAR is an increasing function of the sample size T, and is larger than the AR order of the identified VARMA DGP. We let the lag length be $\lceil \ln T \rceil$, *i.e.* the smallest integer that is greater than $\ln T$, as was suggested by Lütkepohl and Poskitt (1996). The residual obtained from this VAR($\lceil \ln T \rceil$), namely $\hat{\boldsymbol{u}}_t^0$, is a consistent estimate of the true error $\boldsymbol{u}_t$.

The cointegrating vectors in the error correction model are estimated in the first step of the IOLS procedure. We calculate the partial canonical correlations between $\Delta \boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ after controlling for $\Delta \boldsymbol{y}_{t-1}$, ..., $\Delta \boldsymbol{y}_{t-p+1}$ and $\hat{\boldsymbol{u}}_{t-1}^0$, ..., $\hat{\boldsymbol{u}}_{t-q}^0$. The canonical covariates corresponding to the largest $\hat{\rho}$ squared partial canonical correlations are taken as the estimated cointegrating vectors, $\hat{\boldsymbol{\beta}}^0$. $(\hat{\boldsymbol{\beta}}^0)' \boldsymbol{y}_{t-1}$ is commonly referred to as the error correction term. The rest of the parameters are then estimated by the OLS regression of equation (17) with $(\hat{\boldsymbol{\beta}}^0)' \boldsymbol{y}_{t-1}$ and $\hat{\boldsymbol{u}}_{t-1}^0$, ..., $\hat{\boldsymbol{u}}_{t-q}^0$ taken as known, subject to its zero

restrictions in equation (17).

There are a few important issues which should be noted in the OLS estimation of equation (17). First, the zero restrictions imposed by the SCMs on the coefficient matrices $\boldsymbol{\Psi}_i$ and $\boldsymbol{\Theta}_j$ should be taken into account in the estimation, $i = 1, \ldots, p$, $j = 1, \ldots, q$. To put it differently, if some elements of $\boldsymbol{\Psi}_i$ or $\boldsymbol{\Theta}_j$ are restricted to be zero, then the corresponding variables need to be excluded from the OLS estimation.

More importantly, the restrictions on $\boldsymbol{\Phi}_0$ should be reflected in the estimation as well. Recall that $\boldsymbol{\Phi}_0$ is a non-singular matrix with unit diagonal elements for the SCM representation. Consider the OLS estimation of the i-th row of the system equation (17), $i = 1, \ldots, K$. If the ij-element of $\boldsymbol{\Phi}_0$ is non-zero, $j = 1, \ldots, K$ and $j \neq i$, the j-th contemporaneous variable, $\Delta y_{j,t}$ should be put on the right hand side as an explanatory variable. Specifically, the OLS estimation should be conducted using the following equation

$$\begin{aligned}\Delta \boldsymbol{y}_t = \boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}^0)' \boldsymbol{y}_{t-1} + \boldsymbol{\Psi}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Psi}_{p-1} \Delta \boldsymbol{y}_{t-p+1} \\ + (\boldsymbol{I} - \boldsymbol{\Phi}_0) \Delta \boldsymbol{y}_t + \boldsymbol{\Theta}_1 \hat{\boldsymbol{u}}^0_{t-1} + \cdots + \boldsymbol{\Theta}_q \hat{\boldsymbol{u}}^0_{t-q} + \boldsymbol{u}_t,\end{aligned}$$

where $\boldsymbol{u}_t$ is the residual.

The contemporaneous variables $(\boldsymbol{I} - \boldsymbol{\Phi}_0)\Delta \boldsymbol{y}_t$ are not included in the estimation of the cointegrating vectors, because it will not affect the estimated values of $\hat{\boldsymbol{\beta}}^0$. This can be seen by pre-multiplying both sides of equation (17) by $\boldsymbol{\Phi}_0^{-1}$:

$$\begin{aligned}\Delta \boldsymbol{y}_t &= \boldsymbol{\Phi}_0^{-1}\boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_0^{-1}\boldsymbol{\Psi}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_0^{-1}\boldsymbol{\Psi}_{p-1} \Delta \boldsymbol{y}_{t-p+1} \\ &\quad + \boldsymbol{\Phi}_0^{-1}\boldsymbol{u}_t + \boldsymbol{\Phi}_0^{-1}\boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Phi}_0^{-1}\boldsymbol{\Theta}_q \boldsymbol{u}_{t-q} \\ &= \widetilde{\boldsymbol{\alpha}}\boldsymbol{\beta}' \boldsymbol{y}_{t-1} + \widetilde{\boldsymbol{\Psi}}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \widetilde{\boldsymbol{\Psi}}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \widetilde{\boldsymbol{u}}_t + \widetilde{\boldsymbol{\Theta}}_1 \boldsymbol{u}_{t-1} + \cdots + \widetilde{\boldsymbol{\Theta}}_q \boldsymbol{u}_{t-q}, \quad (18)\end{aligned}$$

where $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\Phi}_0^{-1}\boldsymbol{\alpha}$, $\widetilde{\boldsymbol{u}}_t = \boldsymbol{\Phi}_0^{-1}\boldsymbol{u}_t$, $\widetilde{\boldsymbol{\Psi}}_i = \boldsymbol{\Phi}_0^{-1}\boldsymbol{\Psi}_i$, and $\widetilde{\boldsymbol{\Theta}}_j = \boldsymbol{\Phi}_0^{-1}\boldsymbol{\Theta}_j$, $i = 1, \ldots, p-1$, $j = 1, \ldots, q$. Hence equation (18) will give rise to different estimates of the coefficient matrices, but the estimates of interest— $\hat{\boldsymbol{\beta}}^0$ will not change.

The same set of rules applies to each iteration of the OLS estimation hereafter, although we do not state this explicitly in each case. The estimated residual is denoted by $\hat{\boldsymbol{u}}_t^1$. In the subsequent iteration of estimating $\hat{\boldsymbol{\beta}}^1$ and the OLS regression in the form of equation (17), $\hat{\boldsymbol{u}}_t^1$ is used in place of $\hat{\boldsymbol{u}}_t^0$. Formally, suppose that the j-th iteration is evaluated and $\hat{\boldsymbol{u}}_t^j$ is obtained. Let $\hat{\Omega}^j$ be the sample covariance matrix of $\hat{\boldsymbol{u}}_t^j$. The IOLS procedure takes the following steps for an error correction VARMA model.

In the $(j+1)$-th iteration, we first calculate the partial canonical correlation between $\Delta \boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ after controlling for $\Delta \boldsymbol{y}_{t-1}$, ..., $\Delta \boldsymbol{y}_{t-p+1}$ and $\hat{\boldsymbol{u}}_{t-1}^j$, ..., $\hat{\boldsymbol{u}}_{t-q}^j$. The estimated cointegrating vectors $\hat{\boldsymbol{\beta}}^{j+1}$ are formed by the canonical covariates correspond to the largest $\hat{\rho}$ sample squared partial canonical correlations. We then use OLS to estimate the regression model of the following form:

$$\Delta \boldsymbol{y}_t = \boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}^{j+1})' \boldsymbol{y}_{t-1} + \boldsymbol{\Psi}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Psi}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + (\boldsymbol{I} - \boldsymbol{\Phi}_0) \Delta \boldsymbol{y}_t + \boldsymbol{\Theta}_1 \hat{\boldsymbol{u}}_{t-1}^j + \cdots + \boldsymbol{\Theta}_q \hat{\boldsymbol{u}}_{t-q}^j + \boldsymbol{u}_t. \quad (19)$$

Denote the residual estimates obtained from equation (19) by $\hat{\boldsymbol{u}}_t^{j+1}$, and its covariance matrix estimate by $\hat{\Omega}^{j+1}$. If the iterative procedure converges such that $\| \ln |\hat{\boldsymbol{\Omega}}^{j+1}| - \ln |\hat{\boldsymbol{\Omega}}^j| \| < \epsilon$ for some pre-specified constant $\epsilon > 0$, then the OLS estimates of the coefficients in equation (19) are adopted. Otherwise, we should proceed to evaluate the $(j+2)$-th iteration.

The sequence of the residual $\hat{\boldsymbol{u}}_t^j$ is redefined with each iteration j, and therefore there is no guarantee that this iterative process will converge. Kapetanios (2003) points out that iterations of $\hat{\boldsymbol{u}}_t^j$ will converge if this procedure produces a contraction mapping. Hence, he suggests to check the eigenvalues of the Jacobian at each iteration. If any of these eigenvalues are greater than unity, then this signals that the iterative process is unlikely to converge. However, it is difficult to implement this procedure in practice when the dimension of the parameter space is high. Hence, it is necessary to set a pre-specified maximum number of iterations, M_{max}. If the convergence condition $\| \ln |\hat{\boldsymbol{\Omega}}^{j+1}| -$

$\ln|\hat{\boldsymbol{\Omega}}^j|| < \epsilon$ cannot be achieved within M_{max} iterations, there are a few possible solutions to resort to.

Similar to the numerical maximum likelihood methods, good starting values of the parameters are important for convergence of the iterative algorithm. We can perturb the initial estimates of the coefficients using $\hat{\boldsymbol{u}}_t^0$, and repeat the iterative procedure a few times. Alternatively, we can use other estimators as the starting values. To name a few, the Hannan-Rissanen method (Hannan and Rissanen, 1982), the Hannan-Kavalieris procedure (Hannan and Kavalieris, 1984) and the generalized least squares procedure proposed by Koreisha and Pukkila (1990) can all serve this purpose. Kascha (2012) conducts an extensive comparison of these estimators for stationary VARMA DGPs via Monte Carlo simulations. His results suggest that the algorithm of Hannan and Kavalieris (1984) is generally preferable to other algorithms. Hence, if all attempts to initialize the IOLS procedure with good starting values fail and convergence still cannot be achieved, we suggest to use the estimator given by Hannan and Kavalieris (1984) to produce the final estimates.

References

Athanasopoulos, G., Poskitt, D. S. and Vahid, F. (2012), 'Two Canonical VARMA Forms: Scalar Component Models Vis-à-Vis the Echelon Form', *Econometric Reviews* **31**(1), 60–83.

Athanasopoulos, G. and Vahid, F. (2008*a*), 'A Complete VARMA Modelling Methodology Based on Scalar Components', *Journal of Time Series Analysis* **29**(3), 533–554.

Athanasopoulos, G. and Vahid, F. (2008*b*), 'VARMA versus VAR for Macroeconomic Forecasting', *Journal of Business and Economic Statistics* **26**(2), 237–252.

Clements, M. P. and Hendry, D. F. (1993), 'On the Limitations of Comparing Mean Squared Forecast Errors (with discussion)', *Journal of Forecasting* **12**, 617–637.

Diebold, F. X. and Mariano, R. S. (1995), 'Comparing Predictive Accuracy', *Journal of Business and Economic Statistics* **13**, 253–263.

Dufour, J.-M. and Pelletier, D. (2011), 'Practical Methods for Modelling Weak VARMA Processes: Identification, Estimation and Specification with a Macroeconomic Application', *Discussion Paper,* Department of Economics, McGill University, CIREQ and CIRANO .

Engle, R. F. and Granger, C. W. J. (1987), 'Co-integration and Error Correction: Representation, Estimation, and Testing', *Econometrica* **55**, 251–276.

Giacomini, R. and White, H. (2006), 'Tests of Conditional Predictive Ability', *Econometrica* **74**, 1545–1578.

Gonzalo, J. and Pitarakis, J. (1995), 'Specification via Model Selection in Vector Error Correction Models', *Economic Letters* **60**, 321–328.

Granger, C. (1981), 'Some Properties of Time Series Data and Their Use in Econometric Model Specification', *Journal of Econometrics* **16**(1), 121–130.

Hall, A. D., Anderson, H. M. and Granger, C. W. J. (1992), 'A Cointegration Analysis of Treasury Bill Yields', *The Review of Economics and Statistics* **74**(1), 116–126.

Hannan, E. J. and Deistler, M. (1988), *The Statistical Theory of Linear Systems*, John Wiley.

Hannan, E. J. and Kavalieris, L. (1984), 'Multivariate Linear Time Series Models', *Advances in Applied Probability* **16**, 492–561.

Hannan, E. J. and Rissanen, J. (1982), 'Recursive Estimation of Mixed Autoregressive-Moving Average Order', *Biometrika* **69**, 81–94.

Johansen, S. (1988), 'Statistical Analysis of Cointegrating Vectors', *Journal of Economic Dynamics and Control* **12**, 231–254.

Kapetanios, G. (2003), 'A Note on an Iterative Least-squares Estimation Method for ARMA and VARMA Models', *Economics Letters* **79**, 305–312.

Kascha, C. (2012), 'A Comparison of Estimation Methods for Vector Autoregressive Moving Average Models', *Econometric Reviews* **31**(3), 297–324.

Kascha, C. and Trenkler, C. (2014), 'Simple Identification and Specification of Cointegrated VARMA Models', *Journal of Applied Econometrics* . forthcoming.

Koreisha, S. and Pukkila, T. (1990), 'A Generalized Least-squares Approach for Estimation of Autoregressive Moving Average Models', *Journal of Time Series Analysis* **11**(2), 139–151.

Lütkepohl, H. (2005), *The New Introduction to Multiple Time Series Analysis*, Springer-Verlag, Berlin.

Lütkepohl, H. and Claessen, H. (1997), 'Analysis of cointegrated VARMA processes', *Journal of Econometrics* **80**, 223–239.

Lütkepohl, H. and Poskitt, D. S. (1996), 'Specification of Echelon Form VARMA Models', *Journal of Business and Economic Statistics* **14**, 69–79.

Lütkepohl, H. and Poskitt, D. S. (1998), Consistent Estimation of the Number of Cointegration Relations in a Vector Autoregressive Model, Working papers, Humboldt University, Sonderforschungsbereich 373.

Lütkepohl, H. and Saikkonen, P. (1999), 'Order Selection in Testing for the Cointegration Rank of a VAR Process', *Cointegration, Causality and Forecasting. A Festscherift in Honour of Clive W.J. Granger* pp. 168–199.

Poskitt, D. S. (1992), 'Identification of Echelon Canonical Forms for Vector Linear Processes Using Least Squares', *The Annals of Statistics* **20**, 195–215.

Poskitt, D. S. (2000), 'Strongly Consistent Determination of Cointegrating Rank via Canonical Correlations', *Journal of Business and Economic Statistics* **18**, 77–90.

Poskitt, D. S. (2003), 'On the Specification of Cointegrated Autoregressive Moving Average Forecasting Systems', *International Journal of Forecasting* **19**, 503–519.

Stock, J. H. and Watson, M. W. (1988), 'Testing for Common Trends', *Journal of the American Statistical Association* **83**(404), 1097–1107.

Tiao, G. C. and Tsay, R. S. (1989), 'Model Specification in Multivariate Time Series (with discussion)', *Journal of the Royal Statistical Society B* **51**, 157–213.

West, K. D. (1996), 'Asymptotic Inference About Predictive Ability', *Econometrica* **64**, 1067–1084.

Yang, M. and Bewley, R. (1996), 'On Cointegration Tests for VAR Models with Drift', *Economics Letters* **51**(1), 45–50.

Figure 1: Distribution of the estimated cointegrating rank $\hat{\rho}$ when the true cointegrating rank is $\rho_0 = 1$.

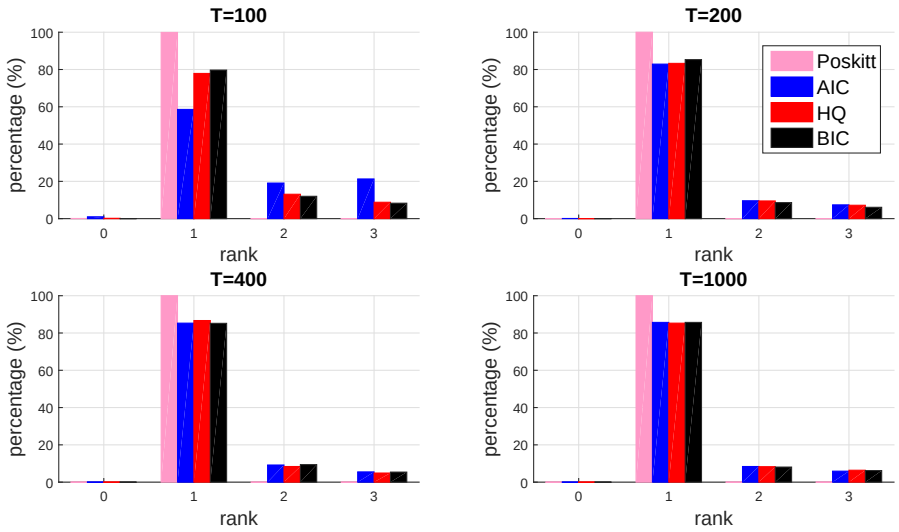

Figure 2: Distribution of the estimated lag length for VARs in levels when the DGP is a cointegrated VARMA(1,1).

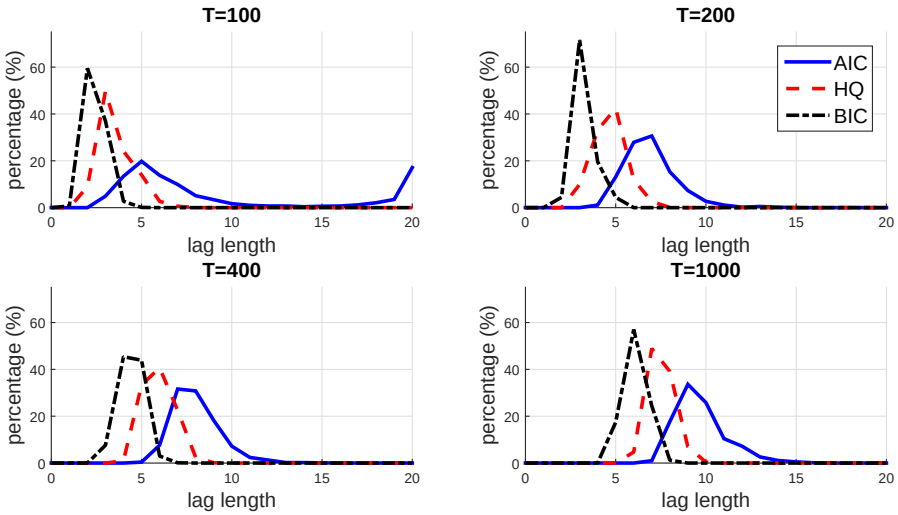

Figure 3: The three interest rate series (%), 1958:12 to 2011:09.

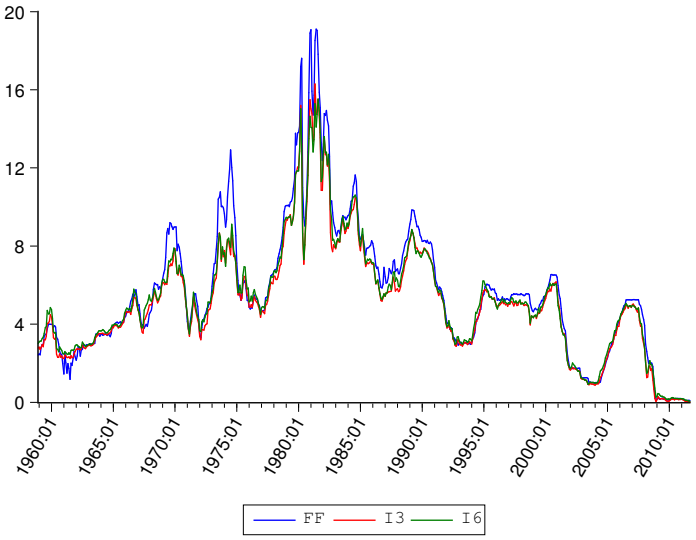

Table 1: Frequency of identified SCMs using the Athanasopoulos and Vahid (2008*a*) SCM identification procedure in a non-stationary setting.

Identified SCMs	(%)			
	$T = 100$	$T = 200$	$T = 400$	$T = 1000$
(1,1)(1,1)(1,1)*	95	95	96	96
(1,2)(1,1)(1,1)	4	3	0	3
(1,2)(1,2)(1,1)	1	1	0	0
(2,2)(1,1)(1,1)	0	0	2	1
Other	0	1	2	0

* The true structure is a cointegrated VARMA(1,1) with three SCMs each of order (1,1).

Table 2: Percentage difference in tr(MSFE) for $\boldsymbol{y}_t$ in levels between the estimated models and the "oracle".

h	VECM							VECM						
	AIC		HQ		BIC		EC-VARMA	AIC		HQ		BIC		EC-VARMA
	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_P$
	$T = 100$							$T = 200$						
1	63.1	61.5	28.8	26.2	29.4	28.5	**16.3**	15.5	16.4	19.6	19.2	21.8	22.2	**10.7**
4	68.0	51.7	21.1	15.0	7.5	6.1	**1.0**	21.9	21.1	16.2	12.6	13.1	11.7	**5.4**
8	76.7	50.4	16.3	12.3	10.5	8.9	**6.7**	17.0	12.5	15.5	8.3	11.0	7.6	**6.0**
12	97.3	59.9	11.8	11.4	**5.5**	9.8	10.0	13.9	7.6	13.0	**5.8**	9.8	7.7	6.7
16	138.2	77.5	14.6	14.3	**8.1**	12.6	11.2	15.6	7.8	15.4	**5.6**	12.2	7.2	5.9
20	175.4	93.1	19.5	17.9	**12.9**	15.6	17.0	14.6	8.9	14.2	**7.4**	12.8	8.3	8.4
24	211.3	98.9	21.8	20.8	**16.3**	18.0	18.6	15.2	10.7	14.3	10.0	14.2	10.0	**9.7**
	$T = 400$							$T = 1000$						
1	9.0	8.5	11.6	11.1	12.6	12.6	**5.6**	4.3	4.4	1.7	1.9	5.0	5.1	**0.8**
4	9.7	11.2	7.5	8.6	4.8	6.3	**1.8**	4.5	4.8	3.0	3.2	4.0	3.9	**-2.4**
8	9.8	10.1	7.2	8.0	5.5	5.9	**3.2**	2.8	3.6	1.5	2.3	1.4	1.9	**-1.7**
12	7.9	6.6	5.0	4.5	3.1	3.2	**2.4**	-0.4	0.2	-2.0	-0.6	**-2.6**	-1.4	-1.7
16	8.5	5.3	6.9	4.4	4.1	1.8	**1.1**	0.1	0.1	-2.1	-0.4	-2.4	-0.9	**-2.9**
20	8.9	5.2	7.3	4.9	5.2	2.7	**1.6**	-0.6	-1.1	**-2.3**	-1.6	-1.3	-1.1	-2.2
24	8.3	5.6	5.9	5.3	5.5	3.9	**2.6**	-0.2	-0.8	-2.2	-1.9	-0.8	-0.6	**-2.9**

[1] $\hat{\rho}_P$: using the cointegrating rank selected by the extended Poskitt method;
[2] $\hat{\rho}_J$: using the cointegrating rank selected by the Johansen procedure;
[3] Bold entries indicate the lowest tr(MSFE) among the estimated models. Negative entries show an improvement over the "oracle".

Table 3: Percentage difference in GFESM between the estimated models and the "oracle"

h	VECM						EC-VARMA	VECM						EC-VARMA
	AIC		HQ		BIC			AIC		HQ		BIC		
	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_J$	$\hat{\rho}_P$	$\hat{\rho}_P$
	$T = 100$							$T = 200$						
1	185.7	179.3	82.9	75.7	85.3	82.9	**48.1**	43.8	46.6	57.1	56.0	63.7	65.2	**29.4**
4	154.6	145.7	54.0	51.4	66.6	66.3	**46.1**	44.4	44.5	39.9	39.7	50.7	50.7	**21.4**
8	124.8	116.9	36.2	**34.4**	45.9	44.9	37.2	30.4	29.1	25.8	24.8	32.3	31.8	**23.5**
12	106.8	100.7	26.5	**25.6**	34.5	34.6	29.5	23.4	22.6	20.2	19.8	25.2	25.2	**18.6**
16	97.2	88.2	20.3	**20.2**	28.9	28.8	25.7	16.8	16.7	15.1	**14.9**	19.9	20.1	18.6
20	91.5	83.0	16.1	**15.9**	24.0	23.6	24.1	15.4	15.8	13.7	**13.7**	15.3	15.9	15.3
24	87.6	79.6	**12.9**	13.4	20.0	19.9	23.1	14.0	14.3	11.5	**11.1**	13.0	13.0	14.7
	$T = 400$							$T = 1000$						
1	24.8	23.5	32.6	31.6	35.5	35.4	**12.9**	9.9	10.3	2.0	2.8	12.5	12.9	**-0.8**
4	23.8	24.9	26.8	27.6	29.2	30.2	**12.6**	16.1	16.5	12.6	12.9	19.8	19.9	**1.7**
8	22.0	22.1	21.5	21.8	20.1	20.0	**9.0**	13.7	14.0	12.2	12.4	14.9	15.0	**1.2**
12	16.8	16.7	15.1	15.3	14.6	14.5	**8.0**	10.0	10.1	9.3	9.5	10.5	10.7	**2.0**
16	13.1	13.1	11.3	11.3	11.6	11.2	**5.5**	8.7	8.7	7.2	7.4	8.0	8.2	**2.0**
20	12.2	12.3	10.3	10.5	9.5	9.2	**6.3**	7.2	7.0	6.6	6.4	7.1	6.8	**0.2**
24	10.5	10.6	9.7	9.6	8.5	8.2	**7.1**	6.6	6.4	5.6	5.3	5.7	5.4	**0.6**

[1] $\hat{\rho}_P$: using the cointegrating rank selected by the extended Poskitt method;
[2] $\hat{\rho}_J$: using the cointegrating rank selected by the Johansen procedure.
[3] Bold entries indicate the lowest GFESM among the estimated models. Negative entries show an improvement over the "oracle".

Table 4: Distributions of the estimated lag lengths for VAR in levels and the cointegrating rank for all 222 estimation windows.

Information criterion	Lag length	(%) Cointegrating rank $\hat{\rho}$		
		0	1	2*
AIC	21	48.2	51.8	
HQ	3			64.9
	9		0.9	11.3
	16			23.0
BIC	2			70.3
	3			29.7
Poskitt's Method				100.0

*$\hat{\rho} = 2$ is the theoretically supported cointegrating rank.

Table 5: Percentage difference in forecast accuracy compared to the benchmark EC-VARMA specifications having imposed theoretically supported cointergating vectors as specified by equation (16).

	Theoretical Cointegration			Data-specified Cointegration			
	VECM			VECM_J			EC-VARMA_P
h	AIC	HQ	BIC	AIC	HQ	BIC	
				tr(MSFE)			
1	56.6	31.5	14.7	59.0^{**} $(^{**},^{**},^{**})$	33.1^{**} $(^{*},^{**},^{**})$	16.3^{*} $(^{*},-,^{*})$	0.9
2	49.6	19.3	7.4	52.8^{**} $(^{**},^{**},^{**})$	20.9^{**} $(^{*},^{**},^{**})$	7.7^{**} $(^{*},-,-)$	2.0
3	48.4	18.4	9.3	51.8^{**} $(^{**},^{**},^{**})$	19.6^{**} $(^{**},^{*},-)$	8.4^{**} $(^{*},-,-)$	3.0
4	40.1	12.1	6.1	43.1^{**} $(^{**},^{**},^{*})$	$13.8^{\dagger}$ $(^{*},-,-)$	5.1 $(^{*},-,-)$	3.8
5	33.4	7.4	4.4	35.5^{**} $(^{**},^{**},^{*})$	9.8 $(^{\dagger},-,-)$	3.8	4.7
6	30.3	5.0	3.8	31.0^{*} $(^{*},^{*},^{\dagger})$	8.4	3.9	5.6
8	23.0	1.8	3.8	21.3	6.8	4.7	6.9
12	17.8	**−1.1**	4.4	11.9	5.6	6.0	7.3
				det(MSFE)			
1	112.4	52.0	17.7	118.9	54.7	28.8	**−4.8**
2	108.4	54.0	20.0	118.5	55.1	32.0	**−12.4**
3	93.0	45.3	18.1	104.2	39.0	21.5	**−20.9**
4	72.4	35.3	16.9	85.4	25.5	14.6	**−22.6**
5	47.6	30.3	23.4	61.7	20.0	16.9	**−25.8**
6	31.0	17.8	17.0	46.2	4.7	5.1	**−28.3**
8	20.0	16.7	17.0	37.2	1.9	1.5	**−35.8**
12	0.0	0.8	16.5	14.5	−26.4	−19.9	**−48.9**
				GFESM			
1	112.4	52.0	17.7	118.9	54.7	28.8	**−4.8**
2	109.0	50.7	12.7	114.2	49.9	17.6	**−7.3**
3	104.5	44.3	7.9	108.7	41.4	9.2	**−8.4**
4	101.5	39.0	6.2	105.2	36.4	6.2	**−6.9**
5	100.4	36.4	4.7	103.5	34.4	4.0	**−5.9**
6	99.9	34.3	3.0	102.6	32.7	1.9	**−5.0**
8	97.9	29.9	2.0	99.7	29.2	1.0	**−3.9**
12	90.3	23.9	0.8	23.8	23.8	−0.0	**−2.2**

The symbol for each tr(MSFE) entry in columns 5-7 represent the Diebold-Mariano test results of equal tr(MSFE) between the EC-VARMA_P and the alternative VECM_J specifications. The null of equality is rejected at **: 1%, *: 5% and †: 10% level of significance against the one sided alternative of lower forecast errors from EC-VARMA_P compared to the alternative. The triplet $(\dot{},\dot{},\dot{})$ below each tr(MSFE) entry represents the results for testing equal MSFE for each ff_t, $i6_t$ and $i3_t$ individually.