Simple Identification and Specification of Cointegrated VARMA Models *

Christian Kascha† Carsten Trenkler‡
University of Zurich *University of Mannheim*

January 22, 2014

Abstract

We bring together some recent advances in the literature on vector autoregressive moving-average models creating a simple specification and estimation strategy for the cointegrated case. We show that in this case with fixed initial values there exists a so-called final moving-average representation. We proof that the specification strategy is consistent. The performance of the proposed method is investigated via a Monte Carlo study and a forecasting exercise for US interest rates. We find that our method performs well relative to alternative approaches for cointegrated series and methods which do not allow for moving-average terms.

JEL classification: C32, C53, E43, E47
Keywords: Cointegration, VARMA Models, Forecasting

1 Introduction

In this paper, we propose a relatively simple specification and estimation strategy for the cointegrated vector autoregressive moving-average (VARMA) model using the estimators given in Yap and Reinsel (1995), Poskitt and Lütkepohl (1995), and Poskitt (2003) and the identified forms proposed by Dufour and Pelletier (2011). We investigate the performance of the proposed methods via a Monte Carlo study and a forecasting exercise for US interest rates and find promising results.

The motivation for looking at this particular model class stems from the well-known theoretical advantages of VARMA models over pure vector-autoregressive (VAR) processes; see e.g. Lütkepohl (2005). In contrast to VAR models, the class of VARMA models is closed under linear transformations. For example, a subset of variables generated by a VAR process

[*] We thank two anonymous referees, the co-editor, Francesco Ravazzolo, Kyusang Yu, Michael Vogt and seminars participants at the Tinbergen Institute Amsterdam, Universities of Aarhus, Bonn, Konstanz, Mannheim, and Zurich as well as session participants at the ISF 2011, Prague, at the ESEM 2011, Oslo, and at the New Developments in Time Series Econometrics conference, Florence, for very helpful comments.

[†] University of Zurich, Chair for Statistics and Empirical Economic Research, Zürichbergstrasse 14, 8032 Zurich, Switzerland; christian.kascha@econ.uzh.ch

[‡] Address of corresponding author: University of Mannheim, Department of Economics, Chair of Empirical Economics, L7, 3-5, 68131 Mannheim, Germany; Phone: +49-621-181-1845; Fax: +49-621-181-1931; trenkler@uni-mannheim.de

is typically generated by a VARMA, not by a VAR process (Lütkepohl, 1984a,b). It is well known that linearized dynamic stochastic general equilibrium (DSGE) models imply that the variables of interest are generated by a finite-order VARMA process. Fernández-Villaverde, Rubio-Ramírez, Sargent and Watson (2007) show formally how DSGE models and VARMA processes are linked. Cooley and Dwyer (1998) claim that modeling macroeconomic time series systematically as pure VARs is not justified by the underlying economic theory. A comparison of structural identification methods using VAR, VARMA and state space representations is provided by Kascha and Mertens (2009).

Existing specification and estimation procedures for cointegrated VARMA models consider sets of parameter restrictions, such as the so-called echelon form or the scalar-component representation, which make sure that the remaining free parameters are identified with respect to the likelihood function. However, while both identified forms can yield representations which are relatively parsimonious, they are often overly complex. Approaches for cointegrated VARMA models that use the (reverse) echelon form can be found in Yap and Reinsel (1995); Lütkepohl and Claessen (1997); Poskitt (2003, 2006) and also Poskitt (2009). The scalar-component representation was originally proposed by Tiao and Tsay (1989) and embedded in a complete estimation procedure by Athanasopoulos and Vahid (2008).

Instead, we extend the final moving-average (FMA) representation of Dufour and Pelletier (2011) to the cointegrated case with fixed initial values. The FMA representation only imposes restrictions on the MA part of the model and, therefore, has a simpler structure than the echelon form. Furthermore, we propose to specify the model using Dufour and Pelletier's (2011) order selection criterion applied to the model estimated in levels. We proof *a.s.* consistency of the estimated orders in this case.

In addition, we compared the proposed approach to the existing approaches via Monte Carlo simulations and via a forecasting exercise. In particular, we apply the methods to the problem of predicting U.S.treasury bill and bond interest rates with different maturities taking cointegration as given. We find rather promising results relative to a variety of different models including a multivariate random walk, the standard vector error correction model (VECM) and approaches based on the echelon form. An investigation of the relative forecasting performances over time shows that our cointegrated VARMA model delivers consistently good forecasts apart from a period stretching from the mid-nineties to 2000.

The rest of the paper is organized as follows. Section 2 discusses our proposals for the identification, specification and estimation of cointegrated VARMA models. In Section 3 we present the results on a Monte Carlo study that investigates how one should implement the proposed procedures and how these compare to alternative methods. Section 4 contains the forecasting study and Section 5 concludes. All programs and data can be found on the homepages of the authors.

2 Cointegrated VARMA models

2.1 Model Framework

The data generating process is formulated here. The assumptions we impose allow us to use the results of Yap and Reinsel (1995), Poskitt and Lütkepohl (1995), Poskitt (2003) and Dufour and Pelletier (2011) in order to construct a reasonably easy and fast strategy for the specification and estimation of cointegrated VARMA models. The considered model for a time series of dimension K, $y_t = (y_{t,1}, \dots, y_{t,K})'$, is

$$A_0 y_t \;=\; \sum_{j=1}^{p} A_j y_{t-j} + \sum_{j=0}^{q} M_j u_{t-j} \text{ for } t = 1, \dots, T, \tag{1}$$

where A_j, $j = 1, \dots, p$, and M_j, $j = 1, \dots, q$ are $K \times K$ parameter matrices with $A_p \neq 0$ and $M_q \neq 0$ and $m := \max\{p, q\}$. The initial values $y_{1-m}, \dots, y_0$ are assumed to be fixed constants.

Let us also define the matrix polynomials $A(z) := A_0 - A_1 z - A_2 z^2 - \ldots - A_p z^p$ and $M(z) := M_0 + M_1 z + \ldots + M_q z^q$, $z \in \mathbb{C}$, with A_0 and M_0 being invertible, and a pair of these by $[A(z), M(z)]$. Using the notation from Poskitt (2006) with minor modifications, we define $\deg[A(z), M(z)]$ as the maximum row degree $\max_{1 \leq k \leq K} \deg_k[A(z), M(z)]$, where $\deg_k[A(z), M(z)]$ denotes the polynomial degree of the kth row of $[A(z), M(z)]$. Then we can define a class of processes by its associated set $\{[A\, M]\}_m := \{[A(z), M(z)] : \deg[A(z), M(z)] = m\}$.

Regarding the error terms we make the following assumption which is equivalent to Assumption A.2 in Poskitt (2003).

Assumption 2.1 *The error term vectors $u_t = (u'_{t,1}, u'_{t,2}, \dots, u'_{t,K})$, $t = 1-m, \dots, 0, 1, \dots, T$, form an independent, identically distributed zero mean white noise sequence with positive definite variance-covariance matrix Σ_u. Furthermore, the moment condition $\mathbf{E}\left(||u_t||^{\delta_1}\right) < \infty$ for some $\delta_1 > 2$, where $||\cdot||$ denotes the Euclidean norm, and growth rate $||u_t|| = O\left((\log t)^{1-\delta_2}\right)$ almost surely (a.s.) for some $0 < \delta_2 < 1$ also hold.*

We make the following two assumptions regarding the polynomials $A(z)$ and $M(z)$:

Assumption 2.2 *$|M(z)| \neq 0$ for $|z| \leq 1$, $z \in \mathbb{C}$, where $|\cdot|$ refers to the determinant.*

Assumption 2.3 *The components in y_t are at most integrated of order one such that $\Delta y_t = y_t - y_{t-1}$ is asymptotically stationary. Moreover, $|A(z)| = a_{st}(z)(1-z)^s$ for $0 < s \leq K$ and $a_{st}(z) \neq 0$ for $|z| \leq 1$, $z \in \mathbb{C}$. The number $r = K - s$ is called the cointegrating rank of the series y_t.*

Hence, the moving-average polynomial is assumed to be invertible. Moreover, it follows from Assumption 2.3 that we can decompose $\Pi := \sum_{j=1}^{p} A_j - A_0$ as $\Pi = \alpha\beta'$, where α and β are $(K \times r)$ matrices with full column rank r. Thus, one can write

$$A_0 \Delta y_t \;=\; \alpha\beta' y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + \sum_{j=0}^{q} M_j u_{t-j} \tag{2}$$

with $\Gamma_j = -(A_{j+1} + \cdots + A_p)$, $t = 1, \dots, T$.

We have not considered a constant term in the specification in (1), mainly for notational convenience. The used estimation methods remain valid provided the constant can be absorbed in the cointegrating relation, i.e. if a constant term in (1) can be expressed as $\mu_0 = -\alpha\rho$, where ρ is of dimension $r \times 1$, such that a linear trend in the variables is ruled out; see Poskitt (2003, Section 2, p. 507) and Yap and Reinsel (1995, Section 6). We follow the approach of Poskitt (2003) and Yap and Reinsel (1995) to accommodate such a constant term by mean-adjusting the data prior to estimation and specification. Hence, we actually apply the methods to $y_t - T^{-1}\sum_{s=1}^{T} y_s$ in the VARMA case. However, the notation in the following will not distinguish between raw and adjusted data. The explicit inclusion of a constant term into the estimation procedure is discussed in Lütkepohl and Claessen (1997).

2.2 Identification

It is well known, that one has to impose certain restrictions on the parameter matrices in order to achieve uniqueness. That is, given a series $(y_t)_{t=1-m}^{T}$, there is generally more than one pair of finite polynomials $[A(z),\, M(z)]$ such that (1) is satisfied. Therefore, one has to restrict the set of considered pairs $[A(z),\, M(z)]$ to a subset such that every process satisfying (1) is represented by exactly one pair in this subset.

Poskitt (2003) proposes a complete modeling strategy using the echelon form which is based on so-called Kronecker indices. Here, we use the much simpler final moving-average (FMA) representation proposed by Dufour and Pelletier (2011) in the context of stationary VARMA models. This representation imposes restrictions on the moving-average polynomial only. More precisely, we consider only polynomials $[A(z),\, M(z)]$, such that

$$M(z) \;=\; m(z)I_K,\; m(z) = 1 + m_1 z + \ldots + m_q z^q. \tag{3}$$

is true and choose among these the pair with the smallest possible orders p, q.[1] As already noted by Dufour and Pelletier (2011), this identification strategy is valid despite $A(z)$ having roots on the unit circle. What is left, is only to show the existence and uniqueness of the FMA form in the non-stationary context with fixed initial values. Analogous to the results in Poskitt (2006), we can show that in this particular case the resulting pair of polynomials does not have to be left-coprime anymore. We assume

Assumption 2.4 *The K-dimensional series $(y_t)_{t=1-m}^{T}$ admits a VARMA representation as in* (1) *with $A_0 = M_0$, $[A(z),\, M(z)] \in \{[A\, M]\}_m$ and fixed initial values $y_{1-m}, \ldots, y_0$.*

The identification of the parameters of the FMA form follows from the observation that any process that satisfies (1) can always be written as

$$y_t = \sum_{s=1}^{t+m-1} \Pi_s y_{t-s} + u_t + n_t,\; t = 1-m, \ldots, T, \tag{4}$$

where it holds, by construction of the sequences $(\Pi_i)_{i=0}^{T+m-1}$ and $(n_t)_{t=1-m}^{T}$, that

$$0 \;=\; \sum_{j=0}^{m} M_j \Pi_{i-j}, \quad i > m \tag{5}$$

$$0 \;=\; \sum_{j=0}^{m} M_j n_{t-j}, \quad t \geq 1. \tag{6}$$

[1] Dufour and Pelletier (2011) also propose another representation that restricts attention to pairs with diagonal moving-average polynomials such as $M(z) = \mathrm{diag}(m_1(z),\, m_2(z),\, \ldots,\, m_K(z))$ where $m_k(z) = 1 + m_{k,1}z + \ldots m_{k,q_k}z^{q_k}$ $k = 1, \ldots, K$ are scalar polynomials. This form delivered results similar to the ones for the FMA form and will therefore not be discussed in the paper.

On the other hand, given a process satisfying (4) and existence of matrices M_0, M_1, ..., M_m such that conditions (5) and (6) are true, the process has a VARMA representation as above. These statements are made precise in the following theorem which is just a restatement of the corresponding theorem in Poskitt (2006).

Theorem 2.1 *The process $(y_t)_{t=1-m}^T$ admits a VARMA representation as in* (1) *with $A_0 = M_0$, A_0 invertible, $[A(z),\, M(z)] \in \{[A\ M]\}_m$ and initial conditions $y_0, \dots, y_{1-m}$ if and only if $(y_t)_{t=1-m}^T$ admits an autoregressive representation*

$$y_t = \sum_{s=1}^{t+m-1} \Pi_s y_{t-s} + u_t + n_t,\ t = 1-m, \dots, T,$$

and there exist matrices $M_0, M_1, \dots, M_m$ which satisfy conditions (5) *and* (6) *and M_0 is invertible.*

Now, one assigns to the autoregressive representation a unique VARMA representation. Because of the properties of the adjoint, $M^{ad}(z)M(z) = |M(z)|$, equations (5) and (6) imply

$$0 = \sum_{j=0}^{\bar{q}} \bar{m}_j \Pi_{i-j}, \quad i > \bar{q} \tag{7}$$

$$0 = \sum_{j=0}^{\bar{q}} \bar{m}_j n_{t-j}, \quad t \geq \bar{q} - m + 1. \tag{8}$$

Here, $|M(z)| =: \bar{m}(z) = \bar{m}_0 + \bar{m}_1 z + \ldots + \bar{m}_{\bar{q}} z^{\bar{q}}$ is a *scalar* polynomial and $\bar{q} = m \cdot K$ is its maximal order.

Because of Theorem 2.1, one can therefore define a pair in final moving-average form as in (3), $[A(z),\, \bar{m}(z)I_K]$, provided that $T \geq \bar{q} - m + 1$ and that the first coefficient is normalized to one. This representation, however, is not the only representation of this form. To achieve uniqueness, we select the representation of the form $[A(z),\, m(z)I_K]$ with the lowest possible degree of the scalar polynomial $m(z)$ such that the first coefficient is normalized to one and (7) *and* (8) are satisfied.

Theorem 2.2 *Assume that the process $(y_t)_{t=1-m}^T$ satisfies Assumption 2.2 and 2.4. Then, for $T \geq \bar{q} - m + 1$, it is always possible to select an observationally equivalent, representation in terms of a pair $[A_0(z),\, m_0(z)I_K]$ with $A_0 = I_K$ and minimal orders p_0 and q_0 commencing from some $t_0 \geq 1 - m$.*

In contrast to the discussion in Dufour and Pelletier (2011) the special feature in the non-stationary case with fixed initial values is that the FMA representation does not need to be left-coprime, in particular the autoregressive and moving-average polynomial can have the same roots. This is a consequence of condition (8) and is not very surprising given the results of Poskitt (2006) on the echelon form representation in the same setting.

If we assume normality and independence, i.e. $u_t \sim i.i.d.N(0, \Sigma_u)$ with Σ_u positive definite, then, under our assumptions, the parameters of the model can be identified via the Gaussian partial likelihood function conditional on the initial observations; see Poskitt (2006, Section 2.2).

The error correction representation

$$y_t = \Pi y_{t-1} + \sum_{i=1}^{p_0-1} \Gamma_j \Delta y_{t-j} + \sum_{j=0}^{q_0} m_{0,j} u_{t-j} \tag{9}$$

with the same initial conditions as above is identified as there exists a one-to-one mapping between this representation and the presentation in levels (cf. Poskitt, 2006, Section 4.1).

2.3 Specification

Dufour and Pelletier (2011) have proposed an information criterion for specifying stationary VARMA models identified via (3). In their setting, the unobserved residuals are first estimated by a long autoregression and then used to fit models of different orders p and q via generalized least squares (GLS). The orders which minimize their information criterion are then chosen. We modify their procedure by replacing the GLS regressions by OLS regressions. We do this in order to be able to apply the results of Huang and Guo (1990) when proving the consistency of the order estimates. The difference between the two variants was mostly irrelevant when they were compared by Monte Carlo simulations (not reported). To be precise, we proceed as follows.

Stage I

Subtract the sample mean from the observations as justified above.

Fit a long VAR regression with h_T lags to the mean-adjusted series as

$$y_t = \sum_{i=1}^{h_T} \Pi_i^{h_T} y_{t-i} + u_t^{h_T}. \tag{10}$$

Denote the estimated residuals from (10) by $\hat{u}_t^{h_T}$.

Stage II

Regress y_t on $\phi_{t-1}(p,q) = [y'_{t-1}, \ldots, y'_{t-p}, \hat{u}_{t-1}^{h_T\prime}, \ldots, \hat{u}_{t-q}^{h_T\prime}]'$, $t = s_T + 1, \ldots T$, imposing the FMA restriction in (3) for all combinations of $p \leq p_T$ and $q \leq q_T$ with $s_T = \max(p_T, q_T) + h_T$ using OLS. Denote the estimate of the corresponding error covariance matrix by $\hat{\Sigma}_T(p,q) = (1/N)\sum_{s_T+1}^{T} z_t(p,q) z'_t(p,q)$, where $z_t(p,q)$ are the OLS residuals and $N = T - s_T$. Compute the information criterion

$$DP(p,q) = \ln|\hat{\Sigma}_T(p,q)| + \dim(\gamma^{(p,q)}) \frac{(\ln N)^{1+\nu}}{N}, \; \nu > 0 \tag{11}$$

where $\dim\left(\gamma^{(p,q)}\right)$ is the dimension of the vector of free parameters of the corresponding VARMA(p,q) model.

Choose the orders by $\widehat{(p,q)}_{IC} = \operatorname{argmin}_{(p,q)} DP(p,q)$, where the minimization is over $p \in \{1, \ldots, p_T\}$, $q \in \{0, 1, \ldots, q_T\}$.

We obtain the following theorem on the consistency of the order estimators.

Theorem 2.3 *If Assumptions 2.1-2.4 hold,* $h_T = [c(\ln T)^a]$ *(the integer part of* $c(\ln T)^a$*) for some* $c > 0$, $a > 1$, *and if* $\max(p_T, q_T) \leq h_T$, *then the orders chosen according to* (11) *converge a.s. to their true values.*

Theorem 2.3 is the counterpart to Dufour and Pelletier (2011, Theorem 5.1), dealing with the stationary VARMA setup, and, to some extent, to Poskitt (2003, Proposition 3.2), referring to cointegrated VARMA models identified via the echelon form. Note, that we can apply the same penalty term $C_T = (\ln N)^{1+\nu}$, $\nu > 0$, as in the stationary VARMA case. However, we use an *i.i.d.* error term assumption in contrast to the strong mixing assumption employed by Dufour and Pelletier (2011). We proceed in this way in order to directly appeal to Poskitt (2003, Proposition 3.2).

The practitioner has to chose values for ν, h_T, p_T, and q_T satisfying the conditions contained in Theorem 2.3. We set $\nu = 0.5$ and $h_T = [(\ln T)^{1.25}]$ according to the results of our own simulations (not reported). The chosen deterministic rule to determine h_T was also applied by Poskitt (2003). For a potential use of information criteria to select h_T see Poskitt (2003, Section 3) and also compare Bauer and Wagner (2005, Corollary 1). Moreover, we set $p_T = q_T = h_T$. Higher orders would lead to near multicollinearity problems due to the fact that the residuals are estimated based on $h_T + 1$ values of y_t. Nevertheless, the maximal lag orders are not very important for the results of the forecasting comparison in Section 4.

2.4 Estimation

The estimation of the model consists of three stages. The first stage is exactly the same as for the specification algorithm. The second stage takes the selected orders as given and estimates the parameters by GLS. Finally, the third stage takes these estimates as a starting point for one iteration of a conditional maximum likelihood iteration step.

Stage I

Again, mean-adjusted data is taken for all stages. For completeness, we restate the equation of the long autoregression here

$$y_t = \sum_{i=1}^{h_T} \Pi_i^{h_T} y_{t-i} + u_t^{h_T} \tag{10'}$$

with estimated residuals $\hat{u}_t^{h_T}$ and covariance estimate $\hat{\Sigma}_u^{h_T} = (T - h_T)^{-1} \sum_{t=h_T+1}^{T} \hat{u}_t^{h_T} \hat{u}_t^{h_T\prime}$.

Stage II

Given orders, p, q, we obtain the estimator of Poskitt and Lütkepohl (1995) and Poskitt (2003) as described in the following.

The cointegrated VARMA model can be conveniently written as

$$\Delta y_t \quad = \quad \Pi' y_{t-1} + [\mathbf{\Gamma}\ \mathbf{M}]\mathbf{Z}_{t-1} + u_t, \tag{12}$$

where $\mathbf{\Gamma} = \text{vec}[\Gamma_1, \dots, \Gamma_k]$, $\mathbf{M} = \text{vec}[M_1, \dots, M_q]$ and $\mathbf{Z}_{t-1} = [\Delta y'_{t-1}, \dots, \Delta y'_{t-k}, u'_{t-1}, \dots, u'_{t-q}]'$. Let $\mathbf{Z}_t^{h_T}$ be the matrix obtained from $\mathbf{Z}_t$ by replacing the u_t by $\hat{u}_t^{h_T}$. Identification restrictions are imposed by defining a suitable restriction matrix, R_1, consisting of zeros and ones such that the vector of free parameters γ_1 relates to the vector of total parameters as $\mathrm{vec}([\Pi\,\mathbf{\Gamma}\,\mathbf{M}]) = R_1\gamma_1$.

Equipped with these definitions, one can write

$$\begin{aligned}\Delta y_t &= \left(y_{t-1}' \otimes I_K,\ \mathbf{Z}_{t-1}^{h_T\,\prime} \otimes I_K\right)\mathrm{vec}([\Pi\,\mathbf{\Gamma}\,\mathbf{M}]) + u_t \\ &= X_t\gamma_1 + u_t., \qquad (13)\end{aligned}$$

where $X_t := \left(y_{t-1}' \otimes I_K,\ \mathbf{Z}_{t-1}^{h_T\,\prime} \otimes I_K\right) R_1$.

Then, the feasible GLS estimator is given by

$$\hat{\gamma}_1 = \left(\sum_{t=h_T+m+1}^{T} X_t'(\hat{\Sigma}_u^{h_T})^{-1}X_t\right)^{-1}\sum_{t=h_T+m+1}^{T} X_t'(\hat{\Sigma}_u^{h_T})^{-1}\Delta y_t, \qquad (14)$$

where $m := \max\{p, q\}$. The estimator is strongly consistent given Assumptions 2.1 - 2.4 (Poskitt, 2003, Propositions 4.1 and 4.2).[2] The estimated matrices are denoted by $\hat{\Pi}, \hat{\mathbf{\Gamma}}, \hat{M}$. To exploit the reduced rank structure in $\Pi = \alpha\beta'$, β is normalized such that $\beta = [I_r, \beta^{*\prime}]'$. Then α is estimated as the first r rows of $\hat{\Pi}$ such that

$$\begin{aligned}\hat{\alpha} &= \hat{\Pi}[., 1:r], \qquad (15)\\ \hat{\beta}^* &= \left(\hat{\alpha}'\left(\hat{M}(1)\hat{\Sigma}_u^{h_T}\hat{M}(1)'\right)^{-1}\hat{\alpha}\right)^{-1} \\ &\times \left(\hat{\alpha}'\left(\hat{M}(1)\hat{\Sigma}_u^{h_T}\hat{M}(1)'\right)^{-1}\hat{\Pi}[., r+1:K]\right). \qquad (16)\end{aligned}$$

See also Yap and Reinsel (1995, Section 4.2) for further details.

Stage III

The estimates from Stage II are taken as starting values for one iteration of a conditional maximum likelihood estimation procedure (Yap and Reinsel, 1995). Define the vector of free parameters given the cointegration restrictions as $\delta := (\mathrm{vec}((\beta^*)')', \mathrm{vec}(\alpha)', \gamma_2')'$, where γ_2 is the vector of unrestricted elements in $\mathbf{\Gamma}, \mathbf{M}$ which is related to the these matrices by the relation $\mathrm{vec}([\mathbf{\Gamma}\ \mathbf{M}]) = R_2\gamma_2$ and R_2 is another restriction matrix imposing the FMA form. Denote the value of δ at the jth iteration as $\delta^{(j)}$. The elements of the initial vector $\delta^{(0)} = \hat{\delta}$ correspond to (14) - (16). Compute $u_t^{(j)}$, $t = 1, \ldots, T$, and $\Sigma_u^{(j)}$ according to

$$\begin{aligned}\sum_{i=0}^{q} M_i^{(j)}u_{t-i}^{(j)} &= \Delta y_t - \alpha^{(j)}\left(\beta^{(j)}\right)' y_{t-1} - \sum_{i=1}^{p-1}\Gamma_i^{(j)}\Delta y_{t-i}, \qquad (17)\\ \Sigma_u^{(j)} &= \frac{1}{T}\sum_{t}^{T} u_t^{(j)}\left(u_t^{(j)}\right)'. \qquad (18)\end{aligned}$$

[2] Our formulation differs from his because we formulate the models in differences throughout. The procedures yield identical results.

For the calculation, it is assumed $y_t = \Delta y_t = u_t = 0$ for $t \leq 0$. Only $W_t^{(j)} := -\partial u_t^{(j)} / \partial \delta_t'$ is needed for computing one iteration of the proposed Newton-Raphson iteration. It can also can be calculated iteratively as

$$\left(W_t^{(j)}\right)' \quad = \quad \left[(y_{t-1}'H \otimes \alpha),\ (y_{t-1}'\beta \otimes I_K),\ ((\mathbf{Z}_{t-1}^{(j)})' \otimes I_K)R_2\right] - \sum_{i=1}^{q} M_i (W_{t-i}^{(j)})' \qquad (19)$$

where $H' := [0_{((K-r)\times r)},\ I_{K-r}]$, c.f. Yap and Reinsel (1995, eqs. (20) and (21)). The estimate is then updated according to

$$\delta^{(j+1)} - \delta^{(j)} = \left(\sum_{t=1}^{T} W_t^{(j)} \left(\Sigma_u^{(j)}\right)^{-1} (W_t^{(j)})'\right)^{-1} \sum_{t=1}^{T} W_t^{(j)} \left(\Sigma_u^{(j)}\right)^{-1} u_t^{(j)}, \qquad (20)$$

which amounts to a GLS estimation step. The estimates of the residuals and their covariance can be updated according to (17) and (18). The one-step iteration estimator $\delta^{(1)}$ is consistent and fully efficient asymptotically according to Yap and Reinsel (1995, Theorem 2) given the strong consistency of the initial estimator $\hat{\gamma}_1$ in (14).

The advantage of this three-step procedure is that it avoids all the complications associated with iterative, nonlinear estimation. In addition, it greatly facilitates the use of simulation-based procedures, like e.g. the bootstrap; a point mentioned by Dufour and Pelletier (2011) in the context of stationary VARMA models.

For the proofs and the forecasting exercise, we take the cointegrating rank as given. However, one might use the results of Yap and Reinsel (1995) to specify the cointegrating rank at the last two steps of the procedure.

3 Monte Carlo Simulations

We have conducted Monte Carlo simulations to show (a) how to specify h_T and ν and (b) how the chosen identification form compares to other identification forms. It turned out that varying h_T and ν had minor effects. Therefore, we present results on (a) in an online supplement and focus on the results on (b) in the paper. To this end, we set $\nu = 0.5$ and $h_T = [(\ln T)^{1.25}]$ which seem to be a reasonable choice according to our simulations.

3.1 Simulation design

For the Monte Carlo simulations we tried to find a spectrum of diverse data generating processes which possibly favor different identification forms. The simulated DGPs are as follows:

DGP I: The process is taken from the Monte Carlo study in Yap and Reinsel (1995) and is a trivariate ARMA(1,1) model with cointegrating rank, $r = 2$, $\Delta y_t = \Pi y_{t-1} + u_t + M_1 u_{t-1}$, where the initial values are set equal to zero, the errors are *i.i.d.* $N(0, \Sigma)$ and the parameter matrices are

$$\alpha\beta' = \begin{pmatrix} -.398 & .433 \\ .121 & -.340 \\ .103 & .166 \end{pmatrix} \begin{pmatrix} 1 & 0 & -.80 \\ 0 & 1 & -.48 \end{pmatrix}, \; M_1 = -\begin{pmatrix} -0.7 & .0 & .0 \\ .3 & -0.5 & .0 \\ -.2 & .1 & .1 \end{pmatrix}$$

$$\Sigma_u = \begin{pmatrix} 1.0 & .5 & .4 \\ .5 & 1.0 & .7 \\ .4 & .7 & 1.0 \end{pmatrix}. \tag{21}$$

DGP II: The process is taken from Lütkepohl and Claessen (1997) that estimated a VARMA with Kronecker indices $\mathbf{p} = (2, 1, 1, 1)$ on US macroeconomic data.[3] The initial values are set to zero. Thus, the process is given by $A_0 \Delta y_{t-1} = \alpha\beta' y_{t-1} + \Gamma_1 \Delta y_{t-1} + A_0 u_{t-1} + M_1 u_{t-1} + M_2 u_{t-2}$, where $\alpha\beta' := A_1 + A_2 - A_0$ and

$$A_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -0.173 & 1 & 0 & 0 \\ -0.350 & 0 & 1 & 0 \\ 0.205 & 0 & 0 & 1 \end{pmatrix}, \; \Gamma_1 = \begin{pmatrix} 0.497 & 0.123 & -0.548 & -0.679 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_1 = \begin{pmatrix} -0.268 & 0 & 0 & 0 \\ 0.143 & 0.035 & 0.490 & -0.373 \\ 0.115 & 0.164 & 0.550 & -0.442 \\ 0.168 & 0.094 & 0.208 & -0.810 \end{pmatrix}, \; M_2 = \begin{pmatrix} 0.151 & 0.112 & 0.104 & 0.464 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\Sigma_u = 10^{-4} \times \begin{pmatrix} 0.699 & 0.076 & -0.100 & -0.401 \\ 0.075 & 0.872 & 0.258 & 0.215 \\ -0.100 & 0.258 & 0.721 & 0.333 \\ -0.401 & 0.215 & 0.333 & 0.790 \end{pmatrix}, \; \alpha = \begin{pmatrix} 0.013 \\ 0.028 \\ 0.00009 \\ 0.0046 \end{pmatrix}, \; \beta = \begin{pmatrix} 1 \\ -0.343 \\ -16.72 \\ 19.35 \end{pmatrix}. \tag{22}$$

DGP III: The parameter values of the trivariate VARMA(1,1) model in FMA form are taken from the estimation results for an interest rate systems with maturities $(3M, 1Y, 10Y)$. This is one of the systems considered in the forecasting study of Section 4. The initial values are set to the actually observed values. The process is $\Delta y_t = \Pi y_{t-1} + u_t + M_1 u_{t-1}$ with

$$\Pi = \begin{pmatrix} -0.30 & 0.31 \\ 0 & -0.02 \\ 0.14 & -0.13 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1.04 \\ 0 & 1 & -1.14 \end{pmatrix}, \; M_1 = \begin{pmatrix} 0.35 & 0 & 0 \\ 0 & 0.35 & 0 \\ 0 & 0 & 0.35 \end{pmatrix},$$

$$\Sigma_u = \begin{pmatrix} 0.16 & 0.14 & 0.06 \\ 0.14 & 0.16 & 0.08 \\ 0.06 & 0.08 & 0.07 \end{pmatrix}. \tag{23}$$

We consider sample sizes of $T = 50, 100, 200$. The results on $T = 200$ are reported in the online supplement. All simulations are based on $R = 1000$ replications.

Since different identification forms lead to different parameters, we cannot compare estimation accuracy. Instead, we compare the accuracy of the implied impulse response estima-

[3] We slightly modified their estimated model by omitting the intercept term such that the generated processes satisfy our assumptions.

tors. Given the moving-average representation of y_t

$$y_t \;=\; \mathbf{m}_t + \sum_{i=0}^{t+m-1} \Phi_i u_{t-i}, \tag{24}$$

where $\mathbf{m}_t$ contains the influence of the initial values, let $\phi_h = \mathrm{vec}(\Phi_h)$ denote the vector of responses of the system to shocks h periods ago. Accuracy is here measured for each horizon as the sum of squared errors of the components $R^{-1}\sum_{i=1}^{R}(\phi_h - \hat{\phi}_{h,i})'(\phi_h - \hat{\phi}_{h,i})$, where $\hat{\phi}_{h,i}$ is the estimated response and the dependence on an estimation and specification strategy is omitted.

We also assess the forecasting precision of the methods. We compute the traces of the estimated mean squared forecast error matrices. These are estimated for horizon h by

$$\mathrm{tr}\,\left(\frac{1}{R}\sum_{i=1}^{R}(y_{T+h,i} - \hat{y}_{T+h|T,i})(y_{T+h,i} - \hat{y}_{T+h|T,i})'\right), \tag{25}$$

where $y_{T+h,n}$ is the value of y_t at $T+h$ for the nth replication and $\hat{y}_{T+h|T,n}$ denotes the corresponding h-step ahead forecast at origin T. The dependence on the specific algorithm is again suppressed. The residuals $\hat{u}_t$ are used to compute forecasts recursively according to

$$\hat{y}_{T+h|T} = A_0^{-1}\left(\sum_{j=1}^{p} A_j \hat{y}_{T+h-j|T} + \sum_{j=h}^{q} M_j \hat{u}_{T+h-j}\right), \tag{26}$$

for $h = 1, \ldots, q$. For $h > q$, the forecast is simply $\hat{y}_{T+h|T} = A_0^{-1}\sum_{j=1}^{p} A_j \hat{y}_{T+h-j|T}$.

3.2 Comparison to cointegrated Echelon VARMA models

There are basically two alternative identification methods: the *scalar component methodology* (SCM) (Tiao and Tsay, 1989; Athanasopoulos and Vahid, 2008) and the identification via a *(reverse) echelon form* as proposed by Lütkepohl and Claessen (1997) for cointegrated time series. Unfortunately, the SCM method cannot be automatized and thus is not comparable in a large scale simulation study, see Athanasopoulos, Poskitt and Vahid (2012). We therefore compare the FMA form only to the echelon representation. The identification via a reversed echelon representation is more complex than the FMA method in that it requires the determination of K integer parameters, the so-called Kronecker indices. In line with the rest of the paper we focus on co-integrated processes.

While the cointegrating rank is assumed to be known, all other specification choices are made data-dependent - in particular we consider the two specification strategies proposed in Poskitt and Lütkepohl (1995) and Poskitt (2003) for the echelon form. They are labeled PL1 and PL2 and determine the Kronecker indices equation by equation by using a selection criterion of the form

$$\Delta_{k,T}(n) = \log \hat{\sigma}^2_{k,T}(n) + C_T n/T, \tag{27}$$

where n is the Kronecker index for equation k, $\hat{\sigma}^2_{k,T}(n)$ is the estimated variance of the corresponding error term and C_T is a penalty term. Since the description of these methods would be too long for the current paper, we recommend Bartel and Lütkepohl (1998) for a concise description. Poskitt and Lütkepohl (1995) proof that both procedures lead to consistent estimators of the true Kronecker indices, provided the penalty term is chosen suitably.

All parameters of the specification methods for the echelon form described above are chosen according to a Monte Carlo study made by Bartel and Lütkepohl (1998). In particular, we have $h_T = \max\{[\ln T], \hat{p}_{AIC}, 4\}$, $P_T = [h_T/2]$ and $C_T = h_T^2$ and $\hat{p}_{AIC}$ is determined by searching up to a maximum order of $1.5 \cdot (\log T)$. These choices are optimized for the echelon form. For the FMA form, we stick to $h_T = [(\ln T)^{1.25}]$ and $\nu = 0.5$

At times, the identified model might lead to an estimated process with highly explosive autoregressive roots. In that case we set the estimated model to a simple multivariate random walk. That case occurred in less than 9 % of the replications for the PL1 specification strategy when applied to realizations of the DGP II with $T = 50$ observations. Otherwise it occurred typically in less than 2 % of the cases for each specification strategy.

— Figures 1 and 2 about here —

Results: Figures 1 and 2 show for DGPs I to III (rows) the estimated MSE for the impulse responses (first column) and the estimated forecast MSEs (second column) of the method using the FMA identification together with the methods relying on the echelon form (PL1 and PL2). The graphs can be summarized as follows. The choice of identification does not matter much for forecast precision while it affects the precision of the impulse response estimators. For $T = 50$ the FMA method is preferable. For higher sample sizes it depends on the DGP. The dynamics of DGP I are relatively well captured by the PL1 and PL2 methods, while the FMA method seems advantageous for DGP II and slightly advantageous for DGP III. This is somewhat surprising given the structure of the simulated DGPs. A possible conjecture is that DGP I has a rich moving-average structure leading to very high orders for the FMA representation. In sum, no method seems to be clearly advantageous in the present context though the results suggest that for small sizes the FMA identification is recommendable.

4 Forecasting Study

We show that our modeling strategy is potentially interesting by applying it in a prediction exercise for US interest rates and comparing the resulting forecasts to those of a set of benchmark models.

The application of the cointegrated VARMA model is motivated by the widespread use of the cointegration approach for the analysis of the term structure of interest rates, see e.g. Campbell and Shiller (1987), Shea (1992), Hassler and Wolters (2001). However, results on the performance of (cointegrated) VARMA models in this context are considerably more sparse. Related to our approach are the papers of Yap and Reinsel (1995), Monfort and Pegoraro (2007) and Feunou (2009).

An alternative approach to time series modeling of the term structure is based on factor models such as the dynamic version of the Nelson-Siegel (DNS) model (Diebold and Li, 2006; Christensen, Diebold and Rudebusch, 2011). These models express the observable interest rates as linear combinations of factors that are often assumed to follow a VAR process. Then, these models also imply a reduced form VARMA representation for the interest rates according to Dufour and Stevanović (2013, Theorem 3.1) but do not embody cointegration restrictions. We do not consider factor models because they would require zero-coupon yield data, like e.g. the Fama-Bliss data which only run until 2003. Furthermore, the forecasting performance of the DNS model was inferior to that of the cointegrated VARMA models on that Fama-Bliss data set.

4.1 Design of Forecast Comparison

We take monthly averages of interest rate data for treasury bills and bonds from the FRED database of the Federal Reserve Bank of St. Louis. The used data are the series `TB3MS`, `TB6MS`, `GS1`, `GS5` and `GS10` with maturities, 3 months, 6 months, 1 year, 5 years and 10 years, respectively. Our vintage starts in 1970:1 and ends in 2012:12 and comprises $T = 516$ data points. Denote by R_{t,m_k} the annualized interest rate for the k-th maturity m_k. Throughout we analyze $y_{t,k} := 100 \ln(1 + R_{t,m_k})$. The data are shown in Figure 3.

— Figure 3 about here —

We compare the forecasting performance of the proposed cointegrated VARMA models using the FMA identification form and the parameter estimates based on the Yap & Reinsel method (EC-VARMA-YR-FMA) to a set of alternative models. We do not discuss the results obtained by using only the initial estimates given in Poskitt and Lütkepohl (1995); Poskitt (2003) because of space constraints and because they appeared to be inferior for some setups.

The primary benchmark is the random walk (RW) model as this model is still regarded quite successful in the literature. Moreover, we consider the VECM as the most closely related cointegrated competitor model. In order to analyze the importance of the specific identification form of a VARMA model, of the moving-average component and of the imposition of cointegration restrictions in more detail, we investigated the forecasting performance of a set of alternative models. These are the cointegrated VARMA model identified via the echelon form using the specification method labeled PL2 as outlined in Section 3 (EC-VARMA-YR-ECH), the level-VARMA model identified via the FMA form (VARMA-FMA) and the level-VAR model. We also investigated the forecasting precision of univariate AR models but excluded them for brevity as they were not systematically superior to the vector models.

The level-VARMA model is specified using the information criterion in (11) as described in Subsection 2.3 and it is estimated setting the cointegration rank r equal to the system dimension K. This approach is equivalent to the estimation procedure proposed in Dufour and Pelletier (2011). This follows from the fact that the iterative step (20) is the same as the third step of Dufour and Pelletier's (2011) procedure if $r = K$. The lag order in levels of the VECM and the level-VAR is specified via the BIC using a maximum lag order of $p_T = [(T/\log T)^{1/2}]$ (Paulsen, 1984; Bauer and Wagner, 2005). The parameters of the VECM are estimated by reduced rank maximum likelihood estimation (Johansen, 1988, 1991, 1995) while we apply the OLS estimator to the level-VAR. Note that the estimation procedures regarding the level-VARMA and VAR models do not impose any restriction on the roots of the autoregressive polynomial. Hence, the estimates could imply a stationary representation of the interest rates.

We chose to select the orders data-dependent for all models in our forecast study. Hence as regards the VARMA and VAR frameworks we compare two modeling strategies rather than two models: one, which allows for nonzero moving-average terms and includes the special case of a pure VAR and one, which exclusively considers the latter case. Moreover, we pre-impose a cointegration rank of $K-1$ on the cointegrated VARMA and vector error correction (VEC) models.

All models are specified and estimated using the data that is available at the forecast origin. Then, forecasts for horizon h are obtained iteratively. The considered forecast horizons are $1, 3, 6$ and 12 months. As the sample expands, all models are re-specified and re-estimated, forecasts are formed and so on - until the end of the available sample is reached. In order to have sufficient observations for estimation, the first forecasts are obtained at $T_s = 200$.

Given estimates of the parameters and innovations, forecasts based on the cointegrated VARMA models are obtained by using the implied VARMA form in levels. Finally, the sample mean, which was subtracted earlier, is added to the forecasts. The latter is also done with respect to the level-VARMA model. The point forecasts based on the RW and the VAR are obtained in a standard way. Similar to the VARMA setup the estimated implied VAR form is used for the VECM to obtain the forecast.

The forecast precision at a certain forecast horizon in terms of an individual series is measured by the (estimated) mean squared prediction errors (MSPEs). The MSPE is defined in a standard way. In the online supplement to this paper we present and discuss the joint forecasting precision, i.e. the forecasting precision with respect to a whole multiple interest rate system. To get a complete picture of the performance of the cointegrated VARMA models vis-a-vis the RW for h-step-ahead forecasts for the k-th series in the system we compute cumulative sums of squared prediction errors defined as

$$\sum_{s=T^s+h}^{t} e^2_{s,RW,k,h} - e^2_{s,\mathcal{M},k,h}, \quad t = T^s + h, \ldots, T, \tag{28}$$

where $\mathcal{M}$ stands for the corresponding model and $\hat{e}_{t,RW,k,h}, \hat{e}_{t,\mathcal{M},k,h}$ are the forecast errors from predicting $y_{t,k}$ based on information up to $t-h$, i.e. $e_{t,\mathcal{M},k,h} = y_{t,k} - \hat{y}_{t,k|t-h,\mathcal{M}}$. Ideally, we should see that the above sum steadily increases over time if forecasting method $\mathcal{M}$ is indeed preferable to the RW. We show the results for the cointegrated VARMA model in FMA form and the VECM for the system with maturities of 3 months and 1 year and forecast horizons $h = 1, 6, 12$ in Figure 4. Similar conclusions can be drawn from pictures regarding the other interest rate systems and model approaches.

4.2 Detailed Results

We have considered all bivariate and three-dimensional models that can be built from the five interest rates as well as the full five-dimensional system. We provide representative findings for some of the interest rate systems in Tables 1 and 2.

— Table 1 about here —

Table 1 contains the main results for the RW model, the cointegrated VARMA and the VECM. The table displays the MSPEs series by series for different systems and horizons. On the left column the maturities of the systems are stated; e.g. the first two rows stand for the bivariate system with interest rates for maturities 3 and 6 months. The MSPEs on the four considered forecast horizons are given in the respective columns labeled as 1, 3, 6, and 12. The entries for the RW model are absolute while the entries for the other models are always relative to the corresponding entry for the RW model. For example, the first entry in the first row tells us that the random walk produces a one-step-ahead MSPE of

0.037 which corresponds to $\sqrt{0.037} \simeq 0.19$ percentage points. In the same row, the entry for EC-VARMA-YR-FMA of 0.739 at $h = 1$ tells us that the cointegrated VARMA-FMA model produces one-step-ahead forecasts of the 3-month interest rate that have a MSPE which is roughly 25 % lower than the MSPE of the RW model.

Table 1 shows that the cointegrated models are more advantageous relative to the RW model for the bivariate systems than for the larger systems. Apparently, the cointegrated models can be very advantageous at one-month and three-month horizons while the RW becomes more competitive for longer maturities and longer horizons - at least when individual MSPEs are considered. While the improved relative performance of the RW in case of longer horizons is in contrast to the findings in Diebold and Li (2006) it fits the results of de Pooter (2007) and Mönch (2008). The latter two studies extend the data set used in Diebold and Li (2006) from to 2000 into 2003, a period in which the RW model performs rather well relative to the DNS model for long horizons. The foregoing discussion clearly indicates to analyze the forecasting performance over time in more detail as we do later on.

Comparing the MSPE figures for the cointegrated VARMA model and the VECM, one can see that the VARMA model is generally performing better, sometimes quite clearly. Typically, a VARMA(1,1) model is preferred by the information criterion over pure VAR models while the BIC usually picks two autoregressive lags regarding the VECM. An exception is the system consisting of five variables. Here, the lag selection criterion (11) almost always chooses no moving-average terms and prefers a VAR(2) like the BIC on the VECM setup. Thus, the "VARMA results" are actually results for the pure VECM model when estimated with the algorithm of Yap and Reinsel (1995). Therefore, the comparison for the five-variable system amounts to a comparison of different estimation algorithms for the same model. It turns out that in this case reduced rank regression is preferable to the Yap and Reinsel's (1995) iterative method in terms of the MSPE measure at $h = 1$. However, at the other forecast horizons the approach of Yap and Reinsel (1995) is superior.

Overall, allowing for moving-average terms in cointegrated models is beneficial. There are two main explanations for this finding. On the one hand, the consideration of MA lags generates a higher degree of model generality. On the other hand, it leads to a smaller number of model parameters in our forecasting study with the exception of the five-dimensional system. As regards the two- and three dimensional systems we usually obtain K^2 and $2 \cdot K^2 - 1$ parameters for the cointegrated VARMA and VEC models, respectively.

Comparing the relative forecast performance over time, Figure 4 of course mirrors the results of Table 1 as the table contains the end-of-sample results. Moreover, the forecasting advantage of the cointegrated VARMA and VEC models is relatively consistent through time. This can be concluded although there are, on the one hand, periods in which the cointegrated models do not perform particularly well and, on the other hand, occasional "jumps" when the cointegrated models perform much better than the RW model. However, these jumps do not appear to dominate the results in Table 1. The outperformance of the cointegrated VARMA over the VEC models is also clearly visible.

There a three periods of particular interest. First, there is a period roughly from the mid-nineties to 2000 when the RW model performed better than the cointegrated models. Interestingly, a similar finding is also obtained by de Pooter, Ravazzolo and van Dijk (2010) in a different context. Note that the level-VAR and VARMA models are also outperformed by the RW during that period. Second, the cointegrated models clearly perform better than the RW in the period between 2001 and 2003. Third, the curves are rather flat since the financial crisis. This results from the low interest rates observed during the recent past. As a consequence, the MSPEs on all models are rather small in magnitude and quite similar to each other. Note that the VARMA model has produced negative interest rate forecasts during the recent years. However, the occurrence is mainly restricted to the 3-month interest rate and the most negative predicted value for this rate is never below -0.35 percentage points.

— Table 2 about here —

Table 2 contains the results on the other models we have considered, i.e. the cointegrated VARMA model identified via the reverse echelon form (EC-VARMA-YR-ECH), the level-VARMA identified via the FMA form (VARMA-FMA) and the level-VAR (VAR). The table is structured like Table 1. Again, the entries show the MSPE values relative to those of the RW model.

Overall, the FMA form outperforms the echelon form in terms of single series' forecast performance. Sometimes the outperformance is quite clear. To give an example, the gain in forecasting precision can amount to roughly 25 percentage points at h=1 for the 3 month interest rate in the bivariate systems (3M, 5Y) and (3M, 10Y).

One potential explanation for the superior performance of FMA form is its smaller number of model parameters compared to the echelon form. The PL2 specification procedure usually suggests a Kronecker index of 1 for each of the K time series regarding all VARMA systems, including the five-dimensional one for which Kronecker indices of 3 are preferred in a few cases. A set of Kronecker indices equal to one leads to VARMA(1,1) systems in which the A_1 and M_1 parameter matrices are not restricted by the reverse echelon form. Hence, the cointegrated VARMA-ECH model contains $2 \cdot K^2 - 1$ parameters in contrast to the FMA form which just has K^2 parameters as described above. Hence, the reduced estimation uncertainty associated with the FMA may have translated into smaller MSPEs, at least in terms of the two- and three-dimensional systems.

Comparing the results of the cointegrated VARMA and the VECM with those of the level-VARMA and level-VAR, respectively, we see that the imposition of $K - 1$ cointegration relations reduces the single series' and systems' MSPEs in almost all of the considered situations, sometimes quite clearly. Hence, the use of multiple time series models in simple level form cannot be recommended for predicting interest rates. A poor performance regarding the VAR has been already documented in the literature, see e.g. Diebold and Li (2006).

Nevertheless, we want to point out the superior forecast performance of the level-VARMA relative to the level-VAR. Hence, the consideration of MA terms is not just beneficial for cointegrated models but also more generally. Therefore, it is worthwhile for applied researchers to consider VARMA-type models rather than to focus on the pure autoregressive framework.

To sum up, the advantage of the cointegrated VARMA-FMA model over the RW is driven by three factors. First, the imposition of cointegration restrictions, second, the consideration of a moving-average component and, third, the use of the FMA identification form.

5 Conclusion

In this paper we tie together some recent advances in the literature on VARMA models creating a relatively simple specification and estimation strategy for the cointegrated case. Our method is based on the final moving average representation that only imposes restrictions on the MA part. In order to show the potential usefulness of our procedure, we applied it in a forecasting exercise for US interest rates and found promising results. In particular, the performance was often superior compared to an approach based on the more complex echelon form.

There are a couple of issues which could be followed up. For example, the appropriate consideration of a linear trend term in the VARMA model would be desirable for many applications. Also, it would be good to augment the model by time-varying conditional variance. Finally, the development of model diagnostic tests appropriate for the cointegrated VARMA case would be of interest.

References

Athanasopoulos G, Poskitt DS, Vahid F. 2012. Two canonical VARMA forms: Scalar component models vis-à-vis the echelon form. *Econometric Reviews* **31**: 60–83.

Athanasopoulos G, Vahid F. 2008. VARMA versus VAR for macroeconomic forecasting. *Journal of Business & Economic Statistics* **26**: 237–252.

Bartel H, Lütkepohl H. 1998. Estimating the Kronecker indices of cointegrated echelon-form VARMA models. *Econometrics Journal* **1**: 76–99.

Bauer D, Wagner M. 2005. Autoregressive approximations of multiple frequency I(1) processes. Economics Series 174, Institute for Advanced Studies.

Campbell JY, Shiller RJ. 1987. Cointegration and tests of present value models. *Journal of Political Economy* **95**: 1062–88.

Christensen JH, Diebold FX, Rudebusch GD. 2011. The affine arbitrage-free class of Nelson-Siegel term structure models. *Journal of Econometrics* **164**: 4–20.

Cooley TF, Dwyer M. 1998. Business cycle analysis without much theory. A look at structural VARs **83**: 57–88.

de Pooter M. 2007. Examining the Nelson-Siegel class of term structure models. Tinbergen Institute Discussion Papers 07-043/4, Tinbergen Institute.

de Pooter M, Ravazzolo F, van Dijk D. 2010. Term structure forecasting using macro factors and forecast combination. *Norges Bank Working Paper 2010/01* .

Diebold FX, Li C. 2006. Forecasting the term structure of government bond yields. *Journal of Econometrics* **130**: 337–364.

Dufour JM, Pelletier D. 2011. Practical methods for modelling weak VARMA processes: Identification, estimation and specification with a macroeconomic application. *Discussion Paper, McGill University, CIREQ and CIRANO* .

Dufour JM, Stevanović D. 2013. Factor-augmented VARMA models with macroeconomic applications. Technical report, McGill University and Carleton University, CIREQ and CIRANO.

Fernández-Villaverde J, Rubio-Ramírez JF, Sargent TJ, Watson MW. 2007. A,B,C's (and D)'s of understanding VARs. *American Economic Review* **97**: 1021–1026.

Feunou B. 2009. A no-arbitrage VARMA term structure model with macroeconomic variables. Technical report, Duke University.

Guo L, Chen HF, Zhang JF. 1989. Consistent order estimation for linear stochastic feedback control systems (CARMA model). *Automatica* **25**: 147–151. ISSN 0005-1098.

Hassler U, Wolters J. 2001. Forecasting money market rates in the unified Germany. In Friedmann R, Knüppel L, Lütkepohl H (eds) *Econometric Studies: A Festschrift in Honour of Joachim Frohn*. LIT, 185–201.

Huang D, Guo L. 1990. Estimation of nonstationary ARMAX models based on the Hannan-Rissanen method. *The Annals of Statistics* **18**: 1729–1756.

Johansen S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12**: 231–254.

Johansen S. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**: 1551–1580.

Johansen S. 1995. *Likelihood-based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Kascha C, Mertens K. 2009. Business cycle analysis and VARMA models. *Journal of Economic Dynamics and Control* **33**: 267–282.

Lai TL, Wei CZ. 1982. Asymptotic properties of projections with applications to stochastic regression problems. *Journal of Multivariate Analysis* **12**: 346–370.

Lütkepohl H. 1984a. Linear aggregation of vector autoregressive moving average processes. *Economics Letters* **14**: 345–350.

Lütkepohl H. 1984b. Linear transformations of vector ARMA processes. *Journal of Econometrics* **26**: 283–293.

Lütkepohl H. 1996. *Handbook of Matrices.* Wiley: Chichester.

Lütkepohl H. 2005. *New Introduction to Multiple Time Series Analysis.* Springer-Verlag: Berlin.

Lütkepohl H, Claessen H. 1997. Analysis of cointegrated VARMA processes. *Journal of Econometrics* **80**: 223–239.

Mönch E. 2008. Forecasting the yield curve in a data-rich environment: A no-arbitrage factor-augmented VAR approach. *Journal of Econometrics* **146**: 26–43.

Monfort A, Pegoraro F. 2007. Switching VARMA term structure models. *Journal of Financial Econometrics* **5**: 105–153.

Nielsen B. 2006. Order determination in general vector autoregressions. In Ho HC, Ing CK, Lai TL (eds) *Time Series and Related Topics: In Memory of Ching-Zong Wei*, volume 52 of *IMS Lecture Notes and Monograph Series.* Institute of Mathematical Statistics, 93–112.

Paulsen J. 1984. Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis* **5**: 115–127.

Poskitt DS. 2003. On the specification of cointegrated autoregressive moving-average forecasting systems. *International Journal of Forecasting* **19**: 503–519.

Poskitt DS. 2006. On the identification and estimation of nonstationary and cointegrated ARMAX systems. *Econometric Theory* **22**: 1138–1175.

Poskitt DS. 2009. Vector autoregressive moving-average identification for macroeconomic modeling: Algorithms and theory. Monash Econometrics and Business Statistics Working Papers 12/09, Monash University, Department of Econometrics and Business Statistics.

Poskitt DS, Lütkepohl H. 1995. Consistent specification of cointegrated autoregressive moving-average systems. SFB 373 Discussion Papers 1995, 54, Humboldt Universität zu Berlin.

Pötscher BM. 1989. Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. *The Annals of Statistics* **17**: 1257–1274.

Shea GS. 1992. Benchmarking the expectations hypothesis of the interest-rate term structure: An analysis of cointegration vectors. *Journal of Business & Economic Statistics* **10**: 347–366.

Tiao GC, Tsay RS. 1989. Model specification in multivariate time series. *Journal of the Royal Statistical Society, B* **51**: 157–213.

Yap SF, Reinsel GC. 1995. Estimation and testing for unit roots in a partially nonstationary vector autoregressive moving average model. *Journal of the American Statistical Association* **90**: 253–267.

A Proofs

Proof of Theorem 2.1:

$\Rightarrow$: Suppose $(y_t)_{t=1-m}^T$ satisfies (1) given initial conditions. One can view the sequence $(u_t)_{t=1-m}^T$ as a solution to (1) viewed as system of equations for the errors and given initial conditions $u_0, \ldots, u_{1-m}$. Then we know that $(u_t)_{t=1-m}^T$ is the sum of a particular solution and the appropriately chosen solution of the corresponding homogeneous system of equations, $u_t = u_t^P + (-n_t)$, say.

Define the sequence $(\Pi_i)_{i \in \mathbb{N}_0}$ by the recursive relations $\Pi_0 = -I_K$ and

$$\begin{aligned} A_i &= \sum_{j=0}^{i} M_j \Pi_{i-j}, \text{ for } i = 1, \ldots, m & (29) \\ 0 &= \sum_{j=0}^{m} M_j \Pi_{i-j}, \text{ for } i > m & (30) \end{aligned}$$

Define now $(u_t^P)_{t=1-m}^T$ by $u_t^P := y_t - \sum_{s=1}^{t+m-1} \Pi_s y_{t-s}$, where $\sum_{s=1}^{0} \Pi_s y_{t-s} := 0$. Then, $(u_t^P)_{t=1-m}^T$ is indeed a particular solution as for $t \geq 1$

$$\begin{aligned} \sum_{j=0}^{p} M_j u_{t-j}^P &= \sum_{j=0}^{m} M_j \left(y_{t-j} - \sum_{s=1}^{t-j+m-1} \Pi_s y_{t-s-j} \right) \\ &= A_0 y_t - \sum_{j=1}^{m} A_j y_{t-j}. & (31) \end{aligned}$$

Further, define $(n_t)_{t=1-m}^T$ by $n_t = u_t^P - u_t$ for $t = 1-m, \ldots, 0$ and $0 = \sum_{i=0}^{m} M_i n_{t-i}$, for $t = 1, \ldots, T$. Since the initial values determine the rest of the series, we also have $u_t = u_t^P - n_t$ for $t \geq 1$.

Therefore, $y_t = \sum_{s=1}^{t+m-1} \Pi_s y_{t-s} + u_t + n_t$ for $t = 1-m, \ldots, 0$.

$\Leftarrow$: Conversely, suppose $(y_t)_{t=1-m}^T$ admits an autoregressive representation as in (4) and there exist $(K \times K)$ matrices M_j $j = 0, \ldots, m$ such that $0 = \sum_{j=0}^{m} M_j \Pi_{i-j}$ for $i > m$ and $0 = \sum_{j=0}^{m} M_j n_{t-j}$, for $t = 1, \ldots, T$.. Then, for $t = 1, \ldots, T$, it holds that

$$\sum_{j=0}^{m} M_j y_{t-j} = \sum_{j=0}^{m} M_j \sum_{s=1}^{t-j+m-1} \Pi_s y_{t-j-s} + \sum_{j=0}^{m} M_j u_{t-j} + \sum_{j=0}^{m} M_j n_{t-j} \quad (32)$$

Moving all terms involving y_t to the left-hand side yields

$$\begin{aligned} -\sum_{v=0}^{t+m-1} \sum_{j=0}^{\min(v,m)} M_j \Pi_{v-j} y_{t-v} &= -\sum_{v=0}^{m} \big(\sum_{j=0}^{v} M_j \Pi_{v-j} \big) y_{t-v} \\ &= A_0 y_t - \sum_{v=1}^{m} A_v y_{t-v} = \sum_{j=0}^{m} M_j u_{t-j} & (33) \end{aligned}$$

where $A_0 := -M_0 \Pi_0 = M_0$ and $A_v := \sum_{j=0}^{v} M_j \Pi_{v-j}$, $v = 1, \ldots, m$.

Proof of Theorem 2.2:

From Theorem 2.1, $(y_t)_{t=1-m}^T$ has a autoregressive representation with the associated series $(\Pi)_{i=0}^{T+m-1}$ and $(n_t)_{t=1-m}^T$. One considers then the set of all polynomials, $m(z) = 1 + m_1 z + \ldots m_q z^q$, for which

$$0 = \sum_{j=0}^{q} m_j \Pi_{i-j} \tag{34}$$

$$0 = \sum_{j=0}^{q} m_j n_{t-j} \tag{35}$$

is true for $i > q$ and $t \geq \underline{t}$ for some $\underline{t}$, $1 - m \leq \underline{t} \leq T$. Denote this set by S. Because of (7) and (8), we know that the (normalized) determinant, $|M(z)|$, satisfies the above conditions with $q = \bar{q}$ and $\underline{t} = \bar{q} - m + 1$. Therefore, S is not empty. Denote one solution to

$$\min_{m(z) \in S} \deg(m(z)), \tag{36}$$

by $m_0(z)$ with degree q_0 and corresponding t_0, where $\deg : S \to \mathbb{N}$ is the function that assigns the degree to every polynomial in S. Suppose, there is another solution of the same degree $m_1(z) = 1 + m_{1,1}z + \ldots + m_{1,q_0} z^{q_0}$. Since both polynomials are of degree q_0, $a = m_{0,q_0}/m_{1,q_0}$ exists and one gets

$$0 = \sum_{j=0}^{q} (m_{0,j} - a m_{1,j}) \Pi_{i-j} \tag{37}$$

$$0 = \sum_{j=0}^{q} (m_{0,j} - a m_{1,j}) n_{t-j} \tag{38}$$

Then, normalization of the first non-zero coefficient of $(m_0(z) - a m_1(z))$ would give a polynomial in S with degree smaller than q_0, a contradiction. Thus $m_0(z)$ is unique.

Given $m_0(z)$, define $A_{0,0} = I_K$ and $A_{0,v} := \sum_{j=0}^{v} m_{0,j} \Pi_{v-j}$, $v = 1, \ldots, q_0$ and p_0 as the minimal number such that $A_{0,v} = 0$ for $v > p_0$.

Then, condition (34) alone would imply left-coprimeness of $[A_0(z), \; m_0(z) I_K]$ but if $(n_t)_{t=1-m}^T \neq 0$ the minimal orders p_0, q_0 might well be above those of the left-coprime solution to (34).

Proof of Theorem 2.3:

Similar to Guo, Chen and Zhang (1989), we proof $(\hat{p}_T, \hat{q}_T) \to (p_0, q_0)$ *a.s.* by showing that the only limit point of $(\hat{p}_T, \hat{q}_T)$ is indeed (p_0, q_0) with probability one, where p_0 and q_0 are the true lag orders. Thus, the convergence of $\hat{p}_T$ and $\hat{q}_T$ follows, which is equivalent to joint convergence. In order to show this, we demonstrate that the events "$(\hat{p}_T, \hat{q}_T)$ has a limit point (p, q) with $p + q > p_0 + q_0$ " (assuming $p \geq p_0, q \geq q_0$) and "$(\hat{p}_T, \hat{q}_T)$ has a limit point (p, q) with $p < p_0$ or $q < q_0$ " both have probability zero.

Following Huang and Guo (1990) we rely on the spectral norm in order to analyze the convergence behaviour of various sample moments; that is, for a $(m \times n)$ matrix A, $||A|| := \sqrt{\lambda_{\max}(A\, A')}$, where $\lambda_{\max}(\cdot)$ denotes the maximal eigenvalue. Lütkepohl (1996, Ch.

8) provides a summary of the properties of this norm. The stochastic order symbols o and O are understood in the context of almost sure convergence.

Case 1: $p \geq p_0, q \geq q_0, p+q > p_0+q_0$

For simplicity, write T instead of N in our lag selection criterion (11). Then

$$DP(p,q) - DP(p_0,q_0) \quad = \quad \ln[|\hat{\Sigma}_T(p,q)|/|\hat{\Sigma}_T(p_0,q_0)|] + c\frac{(\ln T)^{1+v}}{T}, \tag{39}$$

where $c > 0$ is a constant.

We have to show that $DP(p,q) - DP(p_0,q_0)$ has a positive limit for any pair p,q with $p_0 \leq p \leq p_T$, $q_0 \leq q \leq q_T$, and $p+q > p_0+q_0$. Similar to Nielsen (2006, Proof of Theorem 2.5), it is sufficient to show that $T(\hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q)) = O\{g(T)\}$ such that $(\ln T)^{1+v}/g(T) \to \infty$ in this case.

Let us introduce the following notation:

$$\begin{aligned}
\phi_t^0(p,q) &= [y_t', \ldots, y_{t-p+1}', u_t', \ldots u_{t-q+1}']' \\
\phi_t(p,q) &= [y_t', \ldots, y_{t-p+1}', (\hat{u}_t^{h_T})', \ldots, (\hat{u}_{t-q+1}^{h_T})']' \\
Y_T &= [y_1', \ldots, y_T']' \\
U_T &= [u_1', \ldots, u_T']' \\
x_t^0(p,q) &= [(\phi_{t-1}^0(p,q)' \otimes I_K)R]' \\
x_t(p,q) &= [(\phi_{t-1}(p,q)' \otimes I_K)R]' \\
X_T^0(p,q) &= [x_1^0(p,q), \ldots, x_T^0(p,q)]' \\
X_T(p,q) &= [x_1(p,q), \ldots, x_T(p,q)]' \\
\gamma(p,q) &= [\text{vec}(A_1, A_2, \ldots, A_p)', m_1, m_2, \ldots, m_q]',
\end{aligned} \tag{40}$$

where $\gamma(p,q)$ is the $(K^2 \cdot (p+q) \times 1)$ vector of true parameters such that $A_i = 0$ and $m_j = 0$ for $i > p_0, j > q_0$, respectively, and R is implicitly defined such that $\text{vec}[A_1, \ldots, A_p, M_1, \ldots M_q] = R\gamma(p,q)$.

Then, one can write

$$\begin{aligned}
y_t &= \sum_{i=1}^{p} A_i y_{t-i} + u_t + \sum_{i=1}^{q} M_i u_{t-i} \\
&= [A_1, \ldots, A_p, M_1, \ldots M_q]\phi_{t-1}^0(p,q) + u_t \\
&= (\phi_{t-1}^0(p,q)' \otimes I_K)\text{vec}[A_1, \ldots, A_p, M_1, \ldots M_q] + u_t \\
&= x_t^0(p,q)'\gamma(p,q) + u_t
\end{aligned} \tag{41}$$

in order to summarize the model in matrix notation by

$$Y_T = X_T^0(p,q)\gamma(p,q) + U_T = X_T(p,q)\gamma(p,q) + R_T + U_T, \tag{42}$$

where

$$R_T := [X_T^0(p,q) - X_T(p,q)]\gamma(p,q). \tag{43}$$

R_T does not depend on p, q for $p \geq p_0, q \geq q_0$ and can be decomposed as $R_T = [r_0', r_1', \ldots, r_{T-1}']'$, where r_t, $t = 0, 1, \ldots, T-1$, is a $K \times 1$ vector. Let $Z_T(p, q) = [z_1(p, q)', \ldots, z_T(p, q)']'$ be the OLS residuals obtained from regressing Y_T on $X_T(p, q)$, i.e.

$$\begin{aligned} Z_T(p,q) &= Y_T - X_T(p,q)\left[X_T(p,q)'X_T(p,q)\right]^{-1} X_T(p,q)'Y_T \\ &= [R_T + U_T] - X_T(p,q)\left[X_T(p,q)'X_T(p,q)\right]^{-1} X_T(p,q)'[R_T + U_T]. \end{aligned} \quad (44)$$

The estimator of the error covariance matrix Σ_u in dependence on p and q is given by $\hat{\Sigma}_T(p,q) = T^{-1}\sum_{t=1}^{T} z_t(p,q)z_t(p,q)'$. Furthermore, note that $\hat{\Sigma}_T(p_0, q_0) - \hat{\Sigma}_T(p, q)$ is positive semidefinite since $p \geq p_0$ and $q \geq q_0$ in the current setup. Hence, we have

$$\begin{aligned} &\| \hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q) \| = \lambda_{\max}\left(\hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q)\right) \\ &\quad \leq \operatorname{tr}\left(\hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q)\right) = \operatorname{tr}\left(\hat{\Sigma}_T(p_0,q_0)\right) - \operatorname{tr}\left(\hat{\Sigma}_T(p,q)\right) \\ &\quad = T^{-1}[R_T + U_T]'X_T(p,q)\left[X_T(p,q)'X_T(p,q)\right]^{-1} X_T(p,q)'[R_T + U_T] \\ &\qquad - T^{-1}[R_T + U_T]'X_T(p_0,q_0)\left[X_T(p_0,q_0)'X_T(p_0,q_0)\right]^{-1} X_T(p_0,q_0)'[R_T + U_T]. \end{aligned} \quad (45)$$

We have for the terms on the right-hand side (r.h.s.) of the last equality in (45)

$$\begin{aligned} &[R_T + U_T]'X_T(p,q)\left[X_T(p,q)'X_T(p,q)\right]^{-1} X_T(p,q)'[R_T + U_T] \\ &\quad = O\left(|| \left[X_T(p,q)'X_T(p,q)\right]^{-1/2} X_T(p,q)'[R_T + U_T]||^2\right) \\ &\quad = O\left(|| \left[X_T(p,q)'X_T(p,q)\right]^{-1/2} X_T(p,q)'R_T||^2\right) \\ &\qquad + O\left(|| \left[X_T(p,q)'X_T(p,q)\right]^{-1/2} X_T(p,q)'U_T||^2\right), \end{aligned} \quad (46)$$

where the result holds for all $p \geq p_0$ and $q \geq q_0$.

As in Poskitt and Lütkepohl (1995, Proof of Theorem 3.2), we obtain from Lai and Wei (1982, Theorem 3), with a correction mentioned in Pötscher (1989, p. 1268), for any $m = max(p,q)$

$$\begin{aligned} &|| \left[X_T(p,q)'X_T(p,q)\right]^{-1/2} X_T(p,q)'U_T||^2 \\ &\quad = O\left(\max\left\{1, \ln^+\left(\sum_{n=1}^{s}\sum_{t} ||y_{t-n}||^2 + ||\hat{u}_{t-n}^{h_T}||^2\right)\right\}\right) \\ &\quad = O(\ln\ m) + O\left(\ln\left(O\left\{\sum_{t} ||y_t||^2 + ||\hat{u}_t^{h_T}||^2\right\}\right)\right), \end{aligned} \quad (47)$$

where $\ln^+(x)$ denotes the positive part of $\ln(x)$. Moreover, we have that $\sum_t ||y_t||^2 = O(T^g)$ due to Assumption 3.3, where the growth rate is independent of m, see Poskitt and Lütkepohl (1995, Proof of Theorem 3.2, Proof of Lemma 3.1). Therefore, the second term on the r.h.s. of (47) is $O(\ln T)$ for all m. Hence, the left-hand side (l.h.s.) of (47) is $O(\ln T)$ since $m \leq s_{\mathrm{T}} \leq h_T = [c(\ln T)^a]$, $c > 0$, $a > 1$.

Similar to Poskitt and Lütkepohl (1995, Proof of Theorem 3.2) we obtain from a standard result in least squares

$$
\begin{aligned}
|| \left[X_T(p,q)'X_T(p,q)\right]^{-1/2} X_T(p,q)'R_T||^2 &\leq \sum_{t=0}^{T-1}\sum_{i=1}^{K} r_{i,t}^2 \\
&\leq ||\gamma(p,q)||^2 \cdot \sum_{n=1}^{q}\sum_{t=1}^{T} ||u_{t-n} - \hat{u}_{t-n}^{h_T}||^2 = O(\ln T), \qquad (48)
\end{aligned}
$$

where the last line follows from Poskitt (2003, Proposition 3.1) due to Assumption 3.3, our choice of h_T and since $||\gamma(p,q)|| = constant < \infty$ independent of (p,q). Hence, we have $T^{-1}[R_T + U_T]'X_T(p,q)\left[X_T(p,q)'X_T(p,q)\right]^{-1} \times X_T(p,q)'[R_T + U_T] = O(\ln T/T)$.

Using (46 - 48), we have $\parallel \hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q) \parallel = O\left(\ln T/T\right)$ such that $T(\hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q)) = O\{\ln(T)\}$, the desired result, and therefore $DP(p,q) - DP(p_0,q_0) > 0$ *a.s.* for sufficiently large T.

Case 2: (p,q) with $p < p_0$ or $q < q_0$

For (p,q) with $p < p_0$ or $q < q_0$, write

$$
DP(p,q) - DP(p_0,q_0) \;=\; \ln|I_K + (\hat{\Sigma}_T(p,q) - \hat{\Sigma}_T(p_0,q_0))\hat{\Sigma}_T^{-1}(p_0,q_0))| + o(1) \qquad (49)
$$

As in Nielsen (2006, Proof of Theorem 2.4), it suffices to show that $\liminf_{T\to\infty} \lambda_{\max}(\hat{\Sigma}_T(p,q) - \hat{\Sigma}_T(p_0,q_0)) > 0$. To do so, let us introduce some further notation:

$$
\begin{aligned}
\hat{\gamma}_T(p,q) &= \left[X_T'(p,q)X_T(p,q)\right]^{-1} X_T'(p,q)Y_T \qquad (50)\\
&= [\mathrm{vec}(\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_p)', \hat{m}_1, \hat{m}_2, \ldots, \hat{m}_q]'.
\end{aligned}
$$

and, defining $s_p = \max(p,p_0)$ and $s_q = \max(q,q_0)$,

$$
\hat{\gamma}_T^0(p,q) = [\mathrm{vec}(\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_{s_p})', \hat{m}_1, \hat{m}_2, \ldots, \hat{m}_{s_q}]' \qquad (51)
$$

with $\hat{A}_i = 0$ for $i > p$ and $\hat{m}_i = 0$ for $i > q$. Then, we get

$$
\begin{aligned}
Z_T(p,q) &= Y_T - X_T(p,q)\hat{\gamma}_T(p,q) = Y_T - X_T(s_p,s_q)\hat{\gamma}_T^0(p,q)\\
&= U_T + \tilde{X}_T\gamma(s_p,s_q) + X_T(s_p,s_q)\tilde{\gamma}_T(p,q), \qquad (52)
\end{aligned}
$$

where $\gamma(p,q)$ is defined as above in Case 1,

$$
\tilde{X}_T := (X_T^0(s_p,s_q) - X_T(s_p,s_q)), \qquad (53)
$$

and $\tilde{\gamma}_T(p,q) = \gamma(s_p,s_q) - \hat{\gamma}_T^0(p,q)$.

Let $\tilde{x}_t'$ and $x_t'(s_p,s_q)$ be the typical $K \times (pK^2 + q)$ (sub)matrices of the $TK \times (pK^2 + q)$ matrices $\tilde{X}_T$ and $X_T(s_p,s_q)$, respectively, i.e. the partition of $\tilde{X}_T$ and $X_T(s_p,s_q)$ is analogous to $X_T(p,q)$ above. Then, for $p < p_0$ or $q < q_0$, the residual covariance matrix can be written as

$$\begin{aligned}
\hat{\Sigma}_T(p,q) &= \frac{1}{T}\sum_{t=1}^{T} z_t(p,q)z_t'(p,q) \\
&= \frac{1}{T}\sum_{t=1}^{T} x_t'(s_p,s_q)\tilde{\gamma}_T(p,q)\tilde{\gamma}_T'(p,q)x_t(s_p,s_q) \\
&+ \frac{1}{T}\sum_{t=1}^{T} \left(x_t'(s_p,s_q)\tilde{\gamma}_T(p,q)\right)\left(u_t+\tilde{x}_t'\gamma(s_p,s_q)\right)' \qquad (54) \\
&+ \frac{1}{T}\sum_{t=1}^{T} \left(u_t+\tilde{x}_t'\gamma(s_p,s_q)\right)\left(x_t'(s_p,s_q)\tilde{\gamma}_T(p,q)\right)' \\
&+ \frac{1}{T}\sum_{t=1}^{T} \left(u_t+\tilde{x}_t'\gamma(s_p,s_q)\right)\left(u_t+\tilde{x}_t'\gamma(s_p,s_q)\right)' \\
&= D_{1,T} + (D_{2,T}+D_{2,T}') + D_{3,T},
\end{aligned}$$

where $D_{1,T}$, $D_{2,T}$, and $D_{3,T}$ are equal to the square products and cross products in the above equations, respectively.

Similarly, the residual covariance matrix based on the true orders p_0 and q_0 can be expressed by

$$\hat{\Sigma}_T(p_0,q_0) = D_{1,T}^0 + (D_{2,T}^0 + (D_{2,T}^0)') + D_{3,T}^0, \qquad (55)$$

where $D_{1,T}^0$, $D_{2,T}^0$, and $D_{3,T}^0$ are defined analogously to $D_{1,T}$, $D_{2,T}$, and $D_{3,T}$, respectively, by replacing $\gamma(s_p,s_q)$ with $\gamma(p_0,q_0)$. Then,

$$\begin{aligned}
\hat{\Sigma}_T(p,q) - \hat{\Sigma}_T(p_0,q_0) = D_{1,T} &+ \left(D_{2,T}+D_{2,T}'-D_{1,T}^0-D_{2,T}^0-(D_{2,T}^0)'\right) \\
&+ \left(D_{3,T}-D_{3,T}^0\right). \qquad (56)
\end{aligned}$$

It is easily seen that $D_{3,T}$ and $D_{3,T}^0$ both converge to Σ_u *a.s.*. Therefore, the third term in (56) is o(1). We will further show that $D_{2,T}$, $D_{1,T}^0$, and $D_{2,T}^0$ are $o(1)$ and that $\liminf_{T\to\infty}\lambda_{\max}(D_{1,T}) > 0$ while noting that $D_{1,T}$ is p.s.d. by construction, showing $\liminf_{T\to\infty}\lambda_{\max}(\hat{\Sigma}_T(p,q)-\hat{\Sigma}_T(p_0,q_0)) > 0$, the desired result.

$D_{1,T}$**:** Since $D_{1,T}$ is positive semidefinite by construction, it has at least one nonzero eigenvalue if

$$\begin{aligned}
\lambda_{\max}(D_{1,T}) &= \lambda_{\max}\left(\frac{1}{T}\sum_{t=1}^{T} x_t'(s_p,s_q)\tilde{\gamma}_T(p,q)\tilde{\gamma}_T'(p,q)x_t(s_p,s_q)\right) \\
&= \lambda_{\max}\left(\tilde{\Gamma}_T(p,q)\left(\frac{1}{T}\sum_{t=1}^{T}\phi_t(s_p,s_q)\phi_t'(s_p,s_q)\right)\tilde{\Gamma}_T'(p,q)\right) \qquad (57) \\
&\geq \lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\phi_t(s_p,s_q)\phi_t'(s_p,s_q)\right)||\tilde{\Gamma}_T(p,q)||^2, \\
&> 0,
\end{aligned}$$

where $\tilde{\Gamma}_T(p,q) = [\tilde{A}_1, \ldots, \tilde{M}_q]$ is $\tilde{\gamma}(p,q)$ augmented to matrix form. Then, one obtains $\lim_{T\to\infty} \inf T^{-1}\lambda_{\min}(\sum_{t=1}^T \phi_t(s_p,s_q)\phi_t'(s_p,s_q)) > 0$ *a.s.*, according to Poskitt and Lütkepohl (1995, Proof of Theorem 3.2), and $||\tilde{\Gamma}_T(p,q)||^2 = \mathit{constant} > 0$ from Huang and Guo (1990, p. 1753). This gives $\liminf_{T\to\infty} \lambda_{\max}(D_{1,T}) > 0$.

$D^0_{1,T}$: We have

$$\begin{aligned}\tilde{\gamma}_T(p_0,q_0) &= \gamma(p_0,q_0) - \hat{\gamma}^0_T(p_0,q_0) \\ &= -\left[X_T'(p_0,q_0)X_T(p_0,q_0)\right]^{-1} X_T'(p_0,q_0)\left[\tilde{X}_T\gamma(p_0,q_0) + U_T\right]\end{aligned} \tag{58}$$

due to (42), (50), (51), and (53). Therefore,

$$\begin{aligned}&\left\|\frac{1}{T}\sum_{t=1}^T x_t'(p_0,q_0)\tilde{\gamma}_T(p_0,q_0)\tilde{\gamma}_T(p_0,q_0)'x_t(p_0,q_0)\right\| \\ &\quad\le \frac{1}{T}\sum_{t=1}^T ||x_t'(p_0,q_0)\tilde{\gamma}_T(p_0,q_0)\tilde{\gamma}_T(p_0,q_0)'x_t(p_0,q_0)|| \\ &\quad= \frac{1}{T}\sum_{t=1}^T \tilde{\gamma}_T(p_0,q_0)'x_t(p_0,q_0)x_t'(p_0,q_0)\tilde{\gamma}_T(p_0,q_0) \\ &\quad= \frac{1}{T}\tilde{\gamma}_T(p_0,q_0)'\left(X_T'(p_0,q_0)X_T(p_0,q_0)\right)\tilde{\gamma}_T(p_0,q_0),\end{aligned} \tag{59}$$

and, using the above result on $\tilde{\gamma}_T$,

$$\begin{aligned}&= \frac{1}{T}\left[\tilde{X}_T\gamma(p_0,q_0) + U_T\right]' X_T(p_0,q_0)\left[X_T'(p_0,q_0)X_T(p_0,q_0)\right]^{-1} \\ &\quad \times X_T'(p_0,q_0)\left[\tilde{X}_T\gamma(p_0,q_0) + U_T\right] \\ &= \frac{1}{T}O(\ln T),\end{aligned} \tag{60}$$

where the last line follows from (46-48) of the first part of the proof; compare also Huang and Guo (1990, pp. 1754).

$D_{2,T}$: Defining the $(T \times (s_p + s_q)\cdot K)$ matrix $\Phi_T := [\phi_0(s_p,s_q), \ldots, \phi_{T-1}(s_p,s_q)]'$ and let $U^T := [u_1, \ldots, u_T]'$, $X^T := [\tilde{x}_1'\gamma(s_p,s_q), \ldots, \tilde{x}_T'\gamma(s_p,s_q)]'$, then tedious but straightforward calculations lead to

$$\begin{aligned}
&\left\| \frac{1}{T}\sum_{t=1}^{T}(x_t'(s_p,s_q)\tilde{\gamma}_T(p,q))(u_t+\tilde{x}_t'\gamma(s_p,s_q))' \right\| \\
&\quad = \left\| \frac{1}{T}\sum_{t=1}^{T}\tilde{\Gamma}_T(p,q)\phi_t(s_p,s_q)(u_t+\tilde{x}_t'\gamma(s_p,s_q))' \right\| \\
&\quad = \left\| \frac{1}{T}\tilde{\Gamma}_T(p,q)(\Phi_T'\Phi_T)^{1/2}(\Phi_T'\Phi_T)^{-1/2}\sum_{t=1}^{T}\phi_t(s_p,s_q)(u_t+\tilde{x}_t'\gamma(s_p,s_q))' \right\| \\
&\quad \leq \frac{1}{T}||\tilde{\Gamma}_T(p,q)(\Phi_T'\Phi_T)^{1/2}||\;||(\Phi_T'\Phi_T)^{-1/2}\Phi_T'(U^T+X^T)|| \qquad (61) \\
&\quad \leq \left(\frac{1}{T}\tilde{\Gamma}_T(p,q)(\Phi_T'\Phi_T)\tilde{\Gamma}_T'(p,q)\right)^{1/2} \\
&\quad \times \left(\frac{1}{T}(U_T+\tilde{X}_T\gamma)'(\Phi_T\otimes I_K)((\Phi_T'\Phi_T)^{-1}\otimes I_K)(\Phi_T'\otimes I_K)(U_T+\tilde{X}_T\gamma)\right)^{1/2} \\
&\quad = [O(1)]^{1/2}\left[O\left(\frac{1}{T}\ln T\right)\right]^{1/2} = o(1)
\end{aligned}$$

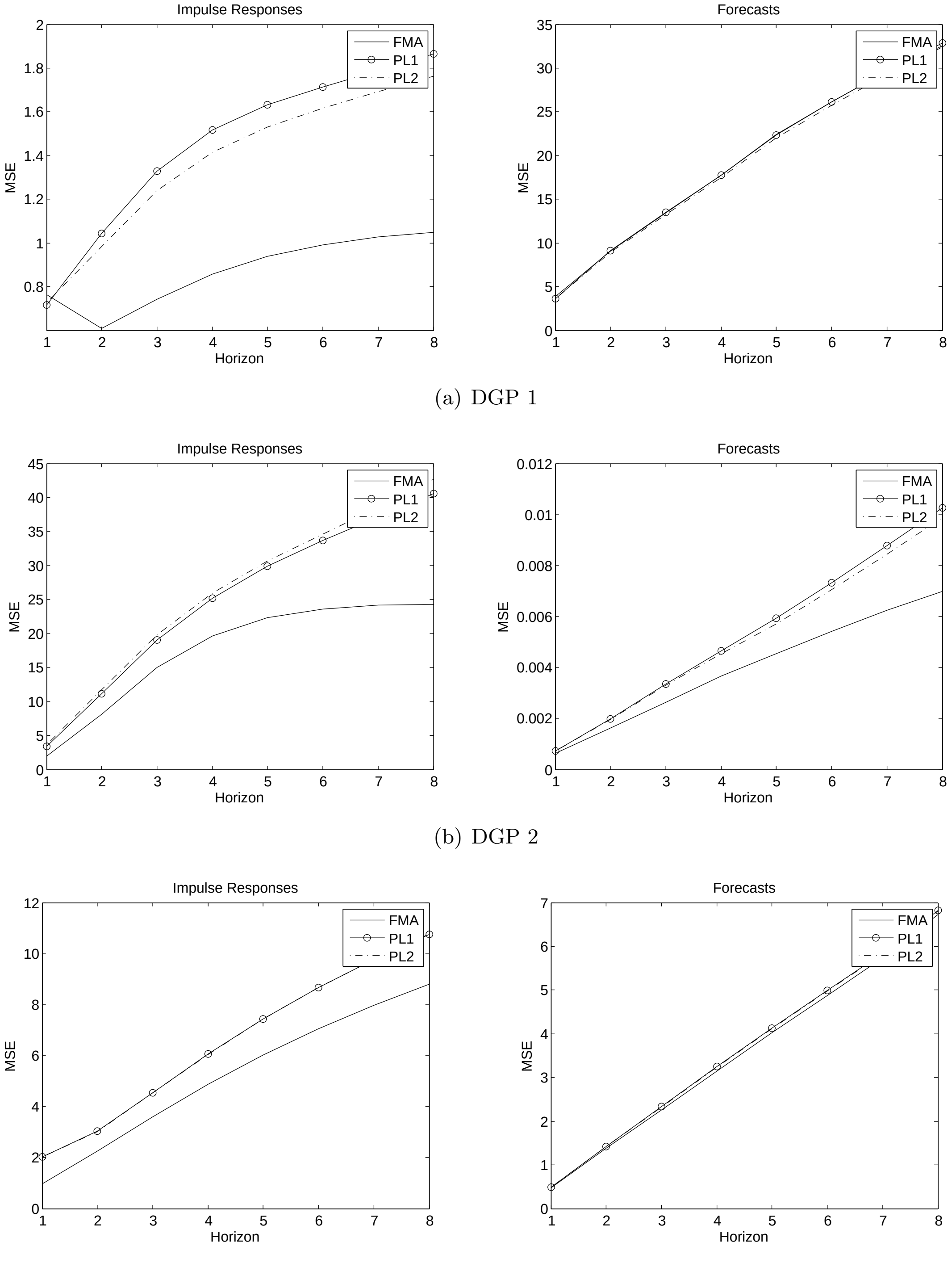
following from the results on $D_{1,T}$ and again from (46-48) of the first part of the proof. Note in this respect that the results in (47) and (48) also hold when using the regressor matrix $\Phi_T \otimes I_K$ appearing in (61). This is due to the fact that the relevant properties of linear projections and OLS do not depend on whether the restricted or unrestricted form of the regressor matrix is used.

$D^0_{2,T}$: Similar to the arguments used for $D_{2,T}$ using arguments identical to those used to evaluate $D^0_{1,T}$, we can write

$$\begin{aligned}
&\left\| \frac{1}{T}\sum_{t=1}^{T}(x_t'\tilde{\gamma}_T(p_0,q_0))(u_t+\tilde{x}_t'\gamma(p_0,q_0)) \right\| \\
&\quad \leq \frac{1}{T}[\tilde{\gamma}_T(p_0,q_0)'(X_T'(p_0,q_0)X_T(p_0,q_0))\tilde{\gamma}_T(p_0,q_0)]^{1/2} \\
&\qquad \times [(U_T+\tilde{X}_T'\gamma(p_0,q_0))'(U_T+\tilde{X}_T'\gamma(p_0,q_0))]^{1/2} \qquad (62) \\
&\quad = \left[O\left(\frac{1}{T}\ln T\right)\right]^{1/2}[O(1)]^{1/2} = o(1)
\end{aligned}$$

noting that $T^{-1}(U_T+\tilde{X}_T'\gamma(p_0,q_0))'(U_T+\tilde{X}_T'\gamma(p_0,q_0)) = T^{-1}U_T'U_T + o(1) = O(1)$ due to the results of Poskitt and Lütkepohl (1995, Proof of Theorem 3.2) and applying Poskitt (2003, Proposition 3.1). This completes the proof.

Figure 1: Estimated Mean Squared Errors for Sample Size $T = 50$

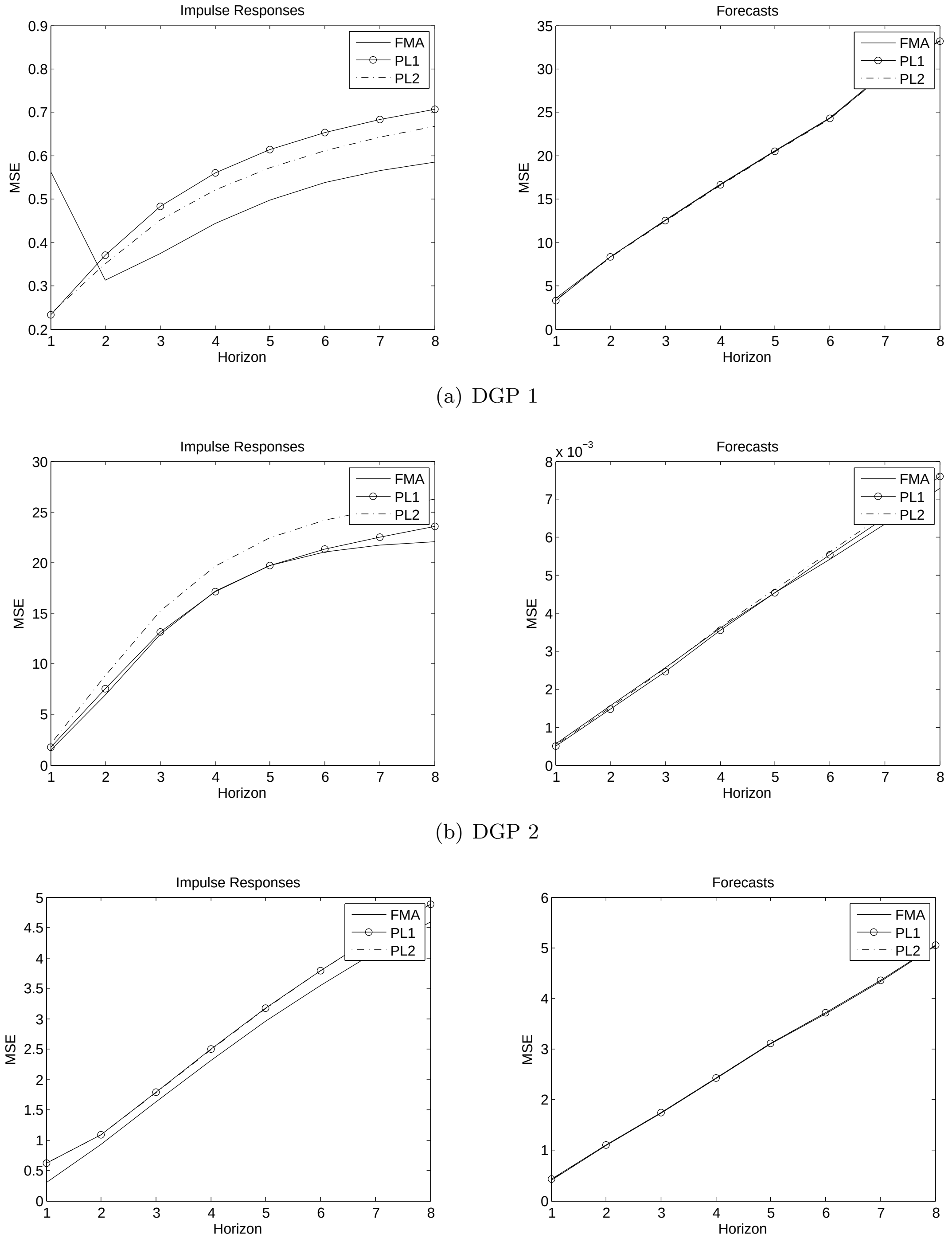


(a) DGP 1

(b) DGP 2

(c) DGP 3

Figure 2: Estimated Mean Squared Errors for Sample Size $T = 100$



(a) DGP 1

(b) DGP 2

(c) DGP 3

Figure 3: US treasury bills and bonds yields. See text for definitions.

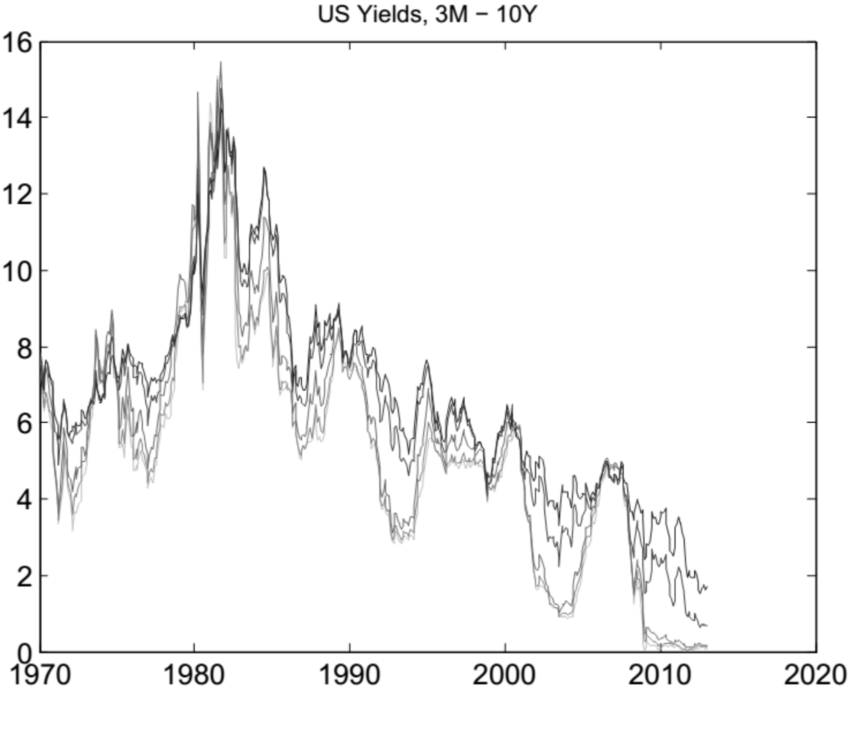
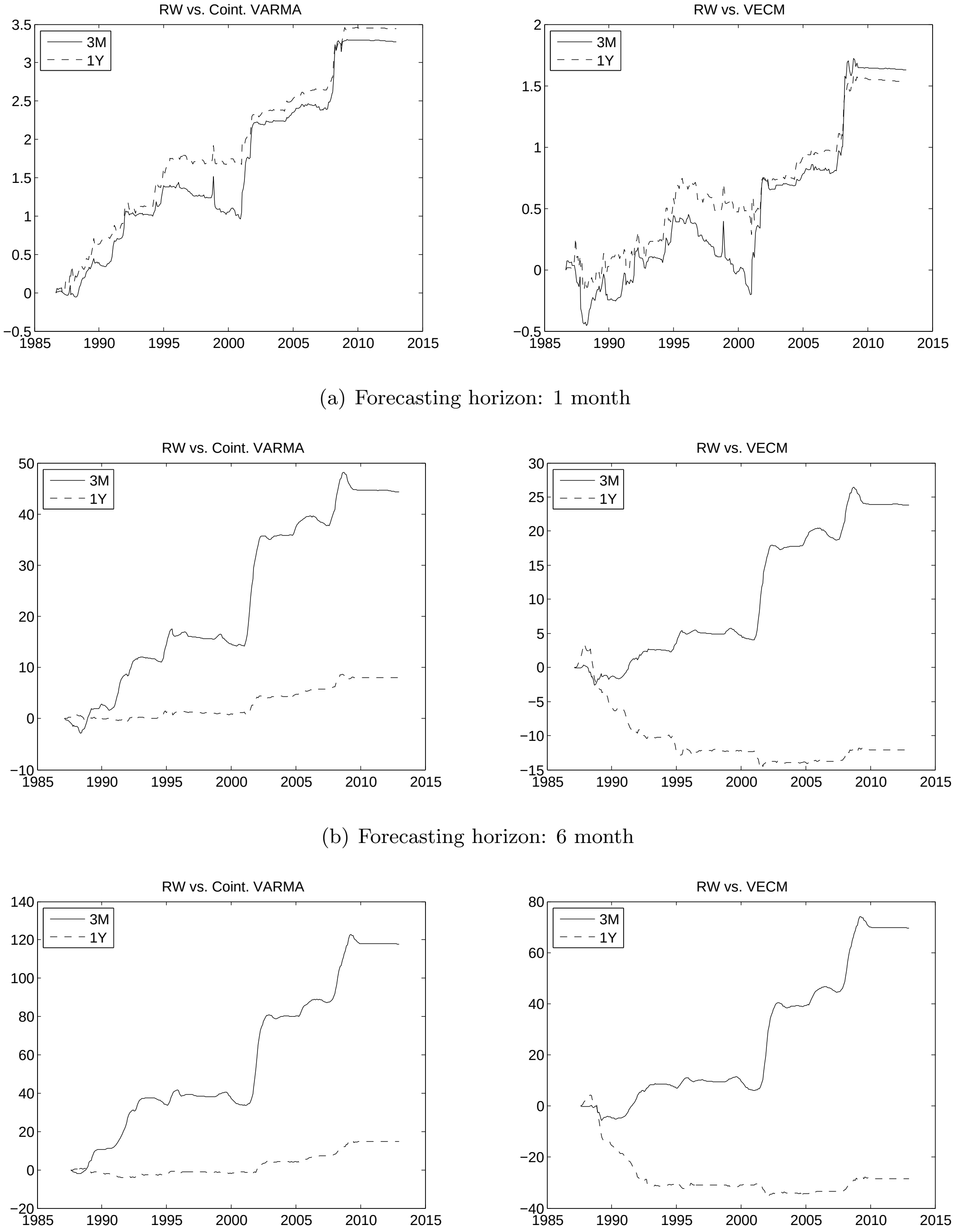


Figure 4: Cumulative squared prediction errors of the cointegrated VARMA model identified via the FMA form and the VECM for different horizons.



(a) Forecasting horizon: 1 month

(b) Forecasting horizon: 6 month

(c) Forecasting horizon: 12 month

Table 1: Comparison of mean squared prediction errors

	RW				EC-VARMA-YR-FMA				VECM			
	1	3	6	12	1	3	6	12	1	3	6	12
3M	0.037	0.203	0.605	1.860	0.749	0.802	0.834	0.884	0.795	0.840	0.885	0.926
6M	0.038	0.209	0.600	1.785	0.787	0.896	0.934	0.959	0.832	0.955	1.002	1.005
3M	0.037	0.203	0.605	1.860	0.722	0.747	0.764	0.791	0.862	0.913	0.874	0.877
1Y	0.049	0.254	0.672	1.832	0.775	0.916	0.960	0.972	0.900	1.089	1.058	1.051
3M	0.037	0.203	0.605	1.860	0.806	0.910	0.914	0.835	1.078	1.069	0.965	0.900
5Y	0.062	0.270	0.570	1.050	0.885	0.979	1.012	1.028	0.947	1.066	1.024	1.062
3M	0.037	0.203	0.605	1.860	0.842	1.004	1.013	0.887	1.106	1.077	0.996	0.905
10Y	0.052	0.207	0.427	0.742	0.910	0.992	1.013	1.004	0.942	1.049	1.020	1.019
3M	0.037	0.203	0.605	1.860	0.900	0.866	0.853	0.848	0.785	0.805	0.891	0.934
6M	0.038	0.209	0.600	1.785	0.951	0.942	0.939	0.916	0.770	0.898	1.001	1.009
1Y	0.049	0.254	0.672	1.832	1.010	1.017	1.025	1.011	0.847	1.028	1.132	1.127
1Y	0.049	0.254	0.672	1.832	0.952	1.162	1.127	1.003	1.089	1.387	1.368	1.105
5Y	0.062	0.270	0.570	1.050	0.930	1.037	1.030	1.033	0.978	1.094	1.126	1.105
10Y	0.052	0.207	0.427	0.742	0.919	0.993	0.981	0.979	0.962	1.026	1.022	1.006
3M	0.037	0.203	0.605	1.860	0.735	0.799	0.847	0.839	0.913	1.095	1.188	1.036
1Y	0.049	0.254	0.672	1.832	0.837	1.029	1.071	1.006	0.990	1.319	1.404	1.204
10Y	0.052	0.207	0.427	0.742	0.914	0.998	1.011	1.004	0.958	1.065	1.092	1.077
3M	0.037	0.203	0.605	1.860	1.249	1.191	1.032	0.870	1.002	1.152	1.114	0.920
6M	0.038	0.209	0.600	1.785	1.229	1.197	1.086	0.934	1.035	1.210	1.199	1.000
1Y	0.049	0.254	0.672	1.832	1.222	1.195	1.108	0.992	1.086	1.249	1.243	1.068
5Y	0.062	0.270	0.570	1.050	1.084	1.055	1.028	1.046	1.039	1.087	1.111	1.106
10Y	0.052	0.207	0.427	0.742	1.026	1.007	0.983	0.985	1.012	1.034	1.037	1.027

Note: The table reports mean squared prediction errors (MSPEs) for systems with different maturities and different models. EC-VARMA-YR-FMA refers to the cointegrated VARMA model identified via the FMA-form and estimated via the method of Yap & Reinsel (1995). The entries for EC-VARMA-YR-FMA and the VECM are relative to the random walk.

Table 2: Comparison of mean squared prediction errors

	EC-VARMA-YR-ECH				VARMA-FMA				VAR			
	1	3	6	12	1	3	6	12	1	3	6	12
3M	0.788	0.810	0.854	0.901	0.786	0.935	1.049	1.124	0.867	1.084	1.244	1.263
6M	0.817	0.918	0.964	0.979	0.872	1.106	1.221	1.252	0.941	1.246	1.405	1.382
3M	0.801	0.786	0.822	0.857	0.800	0.956	1.045	1.089	0.924	1.093	1.063	1.037
1Y	0.823	0.981	1.038	1.044	0.863	1.147	1.285	1.350	0.956	1.242	1.235	1.236
3M	1.063	0.918	0.910	0.862	0.891	1.132	1.195	1.121	1.151	1.267	1.192	1.126
5Y	0.923	0.992	1.038	1.073	0.904	1.049	1.166	1.386	0.960	1.112	1.109	1.256
3M	1.103	0.997	0.974	0.876	0.931	1.201	1.260	1.146	1.168	1.244	1.198	1.122
10Y	0.930	1.009	1.034	1.031	0.924	1.044	1.124	1.298	0.951	1.081	1.082	1.181
3M	0.810	0.800	0.840	0.875	0.968	0.967	0.971	0.960	0.864	1.104	1.357	1.383
6M	0.774	0.874	0.935	0.947	1.055	1.080	1.092	1.058	0.910	1.277	1.530	1.504
1Y	0.813	0.983	1.051	1.056	1.102	1.144	1.175	1.168	0.979	1.391	1.652	1.657
1Y	1.082	1.328	1.280	1.076	1.034	1.360	1.398	1.330	1.196	1.681	1.841	1.694
5Y	0.981	1.095	1.103	1.106	0.971	1.152	1.231	1.411	1.025	1.243	1.427	1.699
10Y	0.965	1.030	1.013	1.012	0.943	1.075	1.135	1.319	0.996	1.143	1.263	1.547
3M	0.888	0.908	0.950	0.914	0.767	0.914	1.032	1.076	0.976	1.284	1.489	1.405
1Y	0.936	1.149	1.184	1.086	0.886	1.161	1.267	1.265	1.054	1.494	1.693	1.592
10Y	0.934	1.032	1.040	1.036	0.929	1.050	1.111	1.245	0.980	1.137	1.239	1.430
3M	1.116	1.276	1.181	0.954	1.378	1.387	1.265	1.123	1.092	1.383	1.456	1.320
6M	1.075	1.308	1.268	1.035	1.371	1.401	1.332	1.201	1.148	1.462	1.564	1.423
1Y	1.069	1.345	1.320	1.102	1.341	1.372	1.336	1.260	1.183	1.476	1.590	1.506
5Y	0.960	1.104	1.132	1.122	1.136	1.149	1.181	1.324	1.088	1.213	1.344	1.557
10Y	0.955	1.034	1.040	1.030	1.065	1.083	1.107	1.233	1.051	1.139	1.231	1.437

Note: The table reports mean squared prediction errors for systems with different maturities and different models relative to the random walk. EC-VARMA-YR-ECH stands for the cointegrated VARMA model identified via the reverse echelon form and estimated via the method of Yap & Reinsel (1995). VARMA-FMA refers to the level-VARMA model identified via the FMA form and estimated via the method of Dufour and Pelletier (2011).