University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN1664-705X(online)

Working Paper No. 33

Cointegrated VARMA Models and Forecasting US Interest Rates

Christian Kascha and Carsten Trenkler

October 2011

Cointegrated VARMA Models and Forecasting US Interest Rates *

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

October 3, 2011

Abstract

We bring together some recent advances in the literature on vector autoregressive moving-average models creating a relatively simple specification and estimation strategy for the cointegrated case. We show that in the cointegrated case with fixed initial values there exists a so-called final moving representation which is usually simpler but not as parsimonious than the usual Echelon form. Furthermore, we proof that our specification strategy is consistent also in the case of cointegrated series. In order to show the potential usefulness of the method, we apply it to US interest rates and find that it generates forecasts superior to methods which do not allow for moving-average terms.

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) models using the estimators given in Yap & Reinsel (1995) and Poskitt (2003) and the identified forms proposed by Dufour & Pelletier (2008). In

*We thank 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; trenkler@uni-mannheim.de

order to show its potential usefulness, we apply the procedure in a forecasting exercise for US interest rates and find promising results.

Existing specification and estimation procedures for cointegrated VARMA models can be found in Yap & Reinsel (1995), Lütkepohl & Claessen (1997), Poskitt (2003, 2006) and also Poskitt (2009). Common to these papers is the use of the reverse "Echelon-Form", a set of parameter restrictions which make sure that the remaining coefficients are identified with respect to the likelihood function. A related but different approach uses so-called "scalar-component" representations originally proposed by Tiao & Tsay (1989) and embedded in a complete estimation procedure by Athanasopoulos & Vahid (2008). While both structures can be quite parsimonious representations of a given process, they can display relatively complex structures. Instead, we extend the simpler identified representations of Dufour & Pelletier (2008) to the cointegrated case with fixed initial values. Furthermore, we propose to specify the model using Dufour & Pelletier's (2008) order selection criteria applied to the model estimated in levels. We proof *a.s.* consistency of the estimated orders in this case. While we believe that our proposed specification and estimation procedure for this class of models stands out because of its simplicity and robustness, this is not to say that our procedure should be preferred to the alternative methods mentioned above under all circumstances. It is likely, that the method of choice depends on the type of data at hand and the sample size. This is therefore an empirical question that goes beyond the scope of this paper.

Finally, 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 quite promising results relative to a multivariate random walk and the standard vector error correction model (VECM). An investigation of the relative forecasting performances over time shows that the VARMA model delivers consistently good forecasts apart from a period stretching from the mid-nineties to 2000.

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 is typically generated by a VARMA, not by a VAR process (Lütkepohl 1984*a*,*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 & Watson (2007) show formally how DSGE models and VARMA processes are linked. Cooley & Dwyer (1998) claim that modeling macroeconomic time series systematically as pure VARs is not justified by the underlying economic theory. A comparsion of structural identification using VAR, VARMA and state space representations is provided by Kascha & Mertens (2009).

Our particular application is part of a vast literature on the term structure of interest rates and serves therefore as an ideal framework in which to compare different modeling strategies. The cointegration approach has become a widespread tool for term structure analysis, following the seminal paper of Campbell & Shiller (1987). If the rational expectation hypothesis of the term structure holds (REHTS) then the spread of two interest rates is stationary. Accordingly, there should exist $K-1$ cointegration relations in a system of K (nonstationary) interest rates. While there is strong empirical evidence for $K-1$ cointegration relations among money market rates and medium-term bond yields, see e.g. Hall, Anderson & Granger (1992), Engsted & Tanggaard (1994), Cuthbertson (1996), Hassler & Wolters (2001), a smaller number of cointegration relations is usually found if long-term interest rates are considered in addition. This has been found e.g. by Shea (1992), Carstensen (2003) and Cavaliere, Rahbek & Taylor (2010).

Starting with Campbell & Shiller (1987), many studies have also analyzed whether the spreads help to predict individual interest rates by classical regression models including the spread or the use of vector autoregressive models. Given the widespread use of cointegration techniques to test for the REHTS, it is, however, surprising that only a few papers apply corresponding multivariate models for cointegration like the VECM for predicting interest rates; see Hall et al. (1992), Hassler & Wolters (2001) and, more recently, Clarida, Sarno, Taylor & Valente (2006)

The results on the forecasting performance of (cointegrated) VARMA models are even more sparse. Using an identified form different from ours, Yap & Reinsel (1995) apply a cointegrated VARMA model to U.S. interest rates but do not evaluate its forecasting performance. An interesting contribution is the one by Feunou (2009). He uses a VARMA model for modeling the whole yield curve imposing no-arbitrage restrictions in a stationary model instead of cointegration restrictions. Another study is provided by Monfort & Pegoraro (2007) using switching VARMA term structure models, again in a stationary context. The applied part of our paper adds to this literature.

The rest of the paper is organized as follows. Section 2 presents the paper's contribution and the application. Section 3 gives details on the proposed methodology. Section 4 concludes. Programs and data can be found on the homepages of the authors.

2 Cointegrated VARMA models

This section summarizes the model framework and the results of the paper. The technical details are given in section 3 and the proofs are provided in the appendix.

This paper mainly assembles and extends elements of the articles of

Yap & Reinsel (1995), Poskitt (2003) and Dufour & Pelletier (2008) 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, $y_t = (y_{t,1}, \ldots, y_{t,K})'$, is

$$y_t \quad = \quad \mu_0 + \sum_{j=1}^{p} A_j y_{t-j} + u_t + \sum_{j=1}^{q} M_j u_{t-j} \text{ for } t = 1, \ldots, T \qquad (1)$$

given fixed initial values $y_0, \ldots, y_{-p+1}$. The error terms u_t are assumed to be *i.i.d.* with mean zero and positive definite covariance matrix, Σ_u, and at least finite second moments (Poskitt 2003, Assumptions A.2, A.3). Let us define the autoregressive and moving-average polynomials by $A(L) = I_K - A_1 L - \ldots - A_p L^p$ and $M(L) = I_K + M_1 L + \ldots + M_q L^q$, respectively, where L denotes the lag operator. $M(L)$ is assumed to be invertible. We are interested in the case in which the process has s unit roots such that $|A(z)| = a_{st}(z)(1-z)^s$ for $0 < s \leq K$ and $|a_{st}(z)| \neq 0$ for $z \leq 1$, where $|\cdot|$ refers to the determinant. Then it is said that the cointegration rank of y_t is $r = K - s$ and we can decompose $\Pi := \sum_{j=1}^{p} A_j - I_K$ as $\Pi = \alpha\beta'$, where α and β are $(n \times r)$ matrices with full column rank r. Furthermore, the constant is assumed to take the form $\mu_0 = -\alpha\rho$ such that a trend in the differences is ruled out and one can write

$$\Delta y_t \quad = \quad \alpha\beta'(y_{t-1} - \rho) + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + u_t + \sum_{j=j}^{q} M_j u_{t-j} \qquad (2)$$

Now, 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), 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 modelling 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 & Pelletier (2008) 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

$$M(L) \quad = \quad m(L) I_K, \; m(L) = 1 + m_1 L + \ldots + m_q L^q. \qquad (3)$$

is true.[1] As already noted by Dufour & Pelletier (2008), this identification strategy is valid despite $A(z)$ having roots on the unit circle. The reason is that the polynomial $M^{-1}(L)A(L)$ can be uniquely related to $M(L)$ and

[1] Dufour & Pelletier (2008) also propose another representation that restricts attention

$A(L)$ (Dufour & Pelletier 2008, Lemma 3.8 and Theorem 3.9). What is left, is only to show the definition 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.

Prior to estimation and specification, we subtract the sample mean from the observations, that is, we actually apply the methods to $y_t - T^{-1}\sum_{s=1}^{T} y_s$ in the VARMA case. However, the notation will not distinguish between raw and adjusted data and we simply write

$$y_t = \sum_{j=1}^{p} A_j y_{t-j} + u_t + \sum_{j=1}^{q} M_j u_{t-j}, \tag{1'}$$

for example. The used estimation methods remain valid, provided the constant can indeed be absorbed in the cointegrating relation; see Yap & Reinsel (1995, section 6.) and Poskitt (2003, section 2, p. 507).

Dufour & Pelletier (2008) also propose 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 which are applied to the cointegrated VARMA model in *levels*.

Having determined the orders p and q, we use the algorithm described in Poskitt (2003) to obtain an initial estimate of the cointegrated model. The estimator basically amounts to an OLS regression in the VECM representation. This estimate is then updated using the algorithm described in Yap & Reinsel (1995) in order to obtain an efficient estimate in a Gaussian setting. The last step is another GLS regression.

The complete procedure for a given cointegrating rank is:

1. Subtract the sample mean from the y_t and estimate a long autoregression using the de-meaned data.

2. Estimate (1') by OLS for different orders p and q imposing the FMA form. The order estimate $(\hat{p}, \hat{q})$ is the pair which minimizes the information criterion in (13).

3. Get a preliminary estimate via Poskitt's method.

---

to pairs with diagonal moving-average polynomials such as

$$M(L) \quad = \quad \bigoplus_{k=1}^{K} m_k(L),\ m_k(L) = 1 + m_{k,1}L + \ldots m_{k,q_k}L^{q_k}$$

where the $m_k(z)$ are scalar polynomials. This form actually delivered results similar to the ones for the FMA form and will therefore not be discussed in the paper.

4. Update the preliminary estimates by the method described by Yap and Reinsel.

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

We show that this modeling strategy is potentially interesting by applying it to a prediction exercise for US interest rates and comparing the resulting forecasts to those of the random walk (RW) model and a VECM that has only autoregressive terms and whose order is chosen by minimizing the BIC. The VECM is estimated via reduced rank regression (Johansen 1988, 1991, 1996).

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 2010:1 and comprises $T = 482$ data points. Denote by R_{t,m_k} the annualized interest rate for the k-th maturity m_k. Throughout we analyze $y_{t,k} := 100 \ln(1 + R_{t,m_k})$.

Both the VAR as well as the VARMA models are specified and estimated using the data that is available at the forecast origin. Then, forecasts for horizon h are obtained iteratively. As the sample expands, both 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 $T_s = 200$. Thus we are left with 282 observations for evaluating 1-month ahead forecasts, for example.

Hence, 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.

Table 1 and Table 2 contain the main results for the RW model, the VECM, the cointegrated VARMA estimated via Poskitt's (2003) initial estimator (VARMA) and the the cointegrated VARMA with estimates updated by one iteration via the algorithm presented in Yap & Reinsel (1995) (VARMA YP). The first table gives the mean square prediction errors (MSPEs) series by series for different systems and horizons. The MSPE is defined in a standard way. The second table gives results for the determinant of the MSPE matrix for different horizons and systems; that is,

$$|MSPE_h| = \left| \frac{1}{T - T_s - h + 1} \sum_{t=T_s}^{T-h} (y_{t+h} - \hat{y}_{t+h|t})(y_{t+h} - \hat{y}_{t+h|t})' \right|,$$

using an obvious notation and omitting the dependence on the specific model and system. The last criterion serves as a criterion to measure joint forecasting precision as we do not want to enter the discussion of how to obtain the

best density forecast. In Table 1, the maturities of the systems are given on the left column; that is, the first two columns stand for the bivariate system with interest rates for maturities 3 and 6 months. The forecast horizons are $1, 3, 6$ and 12 months. Table 2 is structured similarly. For both tables, 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 0.042 which corresponds to $\sqrt{0.042} \simeq 0.205$ percentage points. In the same row, the entry for the VECM at $h = 1$ tells us that this model produces one-step-ahead forecasts of the 3-month interest rate that have a MSPE which is roughly 20 % 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, cointegrated VAR or VARMA models can be very advantegous at one-month and three-month horizons while the RW becomes more competitive for longer horizons and longer maturities - at least when individual MSPEs are considered. The cointegrated models also work better when maturities which are close to each other are grouped together. When comparing the MSPE figures for the VECM and VARMA model (VARMA and VARMA YP) one can see that the VARMA model is generally performing better than the VECM model, sometimes quite clearly. To give an example, the gain in forecasting precision can amount to more than 20% for the bivariate systems at short horizons. Typically, a VARMA(1,1) model is preferred by the information criterion over pure VAR models, while the BIC usually picks two autoregressive lags. An exception is the system consisting of five variables. Here, the lag selection criterion almost always chooses no moving-average terms and thus the "VARMA results" are actually results for the pure VECM model when estimated with the algorithm by Poskitt (2003) or Yap & Reinsel (1995), respectively. Therefore, the comparison for the five-variable system amounts to a comparison of different estimation algorithms for the same model and it turns out that in this case reduced rank regression is largely preferable to the approximative methods in terms of the MSPE measure. Note that the forecasting performance of VARMA and VARMA YP are typically quite similar.

The results in Table 2 largely reflect those in Table 1. That is, the cointegrated models' forecasts are usually more precise than the forecasts generated by the random walk and the VARMA predictions are usually more accurate than the VAR predictions apart from the special case of the five-dimensional system as discussed above. However, in contrast to the single MSPE results, the forecasts generated by the cointegrated models are superior - in terms of joint criterion - to those of the random walk even at longer horizons, in particular for $h = 12$.

To get a complete picture of the performance of the cointegrated 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 as defined as

$$\sum_{s=T^s+h}^{t} e^2_{s,RW,k,h} - e^2_{s,\mathcal{M},k,h} \quad t = T^s + h, \ldots, T \tag{4}$$

where $\mathcal{M}$ stands for either the VECM or the cointegrated VARMA model (VARMA YP) 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 the above sum steadily increasing over time if forecasting method $\mathcal{M}$ is indeed preferable to the RW. The pictures are given in Figure 2 for the system with maturities 3 months and 1 year and forecast horizons $h = 1, 6, 12$.

First, Figure 2 of course mirrors the results of Table 1 as the Table contains the end-of-sample results. Second, the forecasting advantage of both models is relatively consistent through time. While there are occasional "jumps" when the cointegrated models perform much better than the RW model, these jumps do not appear to dominate the results in Table 1. The forecasting advantage of the VARMA model appears however marginally more stable over time. Third, there is also a period roughly from the mid nineties to 2000 when the RW model performed better than the cointegrated multivariate models. Interestingly, a similar finding is also obtained by de Pooter, Ravazzolo & van Dijk (2010) in a different context. In sum, the pictures support the view that the forecasting advantage of both the VECM and VARMA model over the RW is systematic.

3 Methodological Details

3.1 VARMA Modelling

We start discussing the general VARMA process

$$A_0 y_t = \sum_{j=1}^{m} A_j y_{t-j} + M_0 u_t + \sum_{j=1}^{m} M_j u_{t-j}, \text{ for } t = 1, \ldots, T, \tag{5}$$

where, for simplicity, m denotes the maximum of the autoregressive and moving average lag order in this section. Using the notation from Poskitt (2006) with minor modifications, we define $\deg[A(z),\, M(z)]$ as the maximum row degree $\max_{1 \le k \le K} \deg_k[A(z),\, M(z)]$ 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 $\{[A\, M]\}_m := \{[A(z),\, M(z)] | \deg[A(z),\, M(z)] = m\}$.

For the moment we just assume

Assumption 3.1 *The K-dimensional series $(y_t)_{t=1-m}^{T}$ admits a VARMA representation as in* (5) *with $A_0 = M_0$, A_0 invertible, $[A(z),\, M(z)] \in \{[A\, M]\}_m$ and fixed initial values $y_{1-m}, \ldots, y_0$;*

but impose additional restrictions of the general model as needed. The proofs are given in the appendix.

Identification

The identification of the parameters of the FMA form follows from the observation that any process that satisfies (5) can always be written as

$$y_t = \sum_{s=1}^{t+m-1} \Pi_s y_{t-s} + u_t + n_t,\ t = 1-m, \ldots, T, \tag{6}$$

where it holds, by construction of the sequences Π_i and n_t, that

$$0 = \sum_{j=0}^{m} M_j \Pi_{i-j}, \quad i \geq m+1 \tag{7}$$

$$0 = \sum_{j=0}^{m} M_j n_{t-j}, \quad t = 1, \ldots, T. \tag{8}$$

On the other hand, given a process satisfying (6) and existence of matrices M_0, M_1, ..., M_m such that conditions (7) and (8) 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 3.1 *The process $(y_t)_{t=1-m}^{T}$ admits a VARMA representation as in* (5) *with $[A(z),\, M(z)] \in \{[A\ M]\}_m$ and initial conditions $y_0, \ldots, y_{1-m}$ if and only if $(y_t)_{t=1-m}^{T}$ admits an autoregressive representation*

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

in which the conditions (7) *and* (8) *are satisfied.*

Now, one assigns to the autoregressive representation a unique VARMA representation. Although not necessary for the derivations that follow immediately, we assume that $M(z)$ is invertible

Assumption 3.2 $|M(z)| \neq 0$ *for* $|z| \leq 1$.

Because of the properties of the adjoint, $M^{ad}(z)M(z) = |M(z)|$, equations (7) and (8) imply

$$0 = \sum_{j=0}^{\bar{q}} \bar{m}_j \Pi_{i-j}, \quad i \geq \bar{q}+1 \tag{9}$$

$$0 = \sum_{j=0}^{\bar{q}} \bar{m}_j n_{t-j}, \quad t = \bar{q}-m+1,\ldots,T. \tag{10}$$

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

Therefore, one can define a pair in final moving-average form as in (3) $[A(z), \bar{m}(z)I_K]$, provided the stated assumptions and that $T \geq \bar{q} - m + 1$. 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 one and (9) *and* (10) are satisfied.

Theorem 3.2 *Assume that the process $(y_t)_{t=1-m}^{T}$ satisfies Assumptions 3.1 and 3.2. Then for $T \geq \bar{q} - m + 1$ there exists a unique, observationally equivalent, representation in terms of a pair $[A_0(z), m_0(z)I_K]$ with orders p_0 and q_0, respectively, commencing from some $t_0 \geq 1 - m$.*

In contrast to the discussion in Dufour & Pelletier (2008), 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 (10) and is not very surprising given the results of Poskitt (2006) on the Echelon form representation in the same setting.

Now, if we assume normality and independence, i.e. $u_t \sim$ *i.i.d.*$N(0, \Sigma_u)$ with Σ_u positive definite, and under our assumptions, the parameters of the model are identified via the Gaussian partial likelihood function

$$f(y_{t_0}^{T} | y_{1-m}^{t_0-1}, \lambda)$$

where $y_{t_0}^{T} = (y_{t_0}', \ldots, y_T')'$, $y_{1-m}^{t_0-1} = (y_{1-m}', \ldots, y_{t_0-1}')'$ and λ being the parameter vector of the final moving average form. This just follows from Poskitt (2006, section 2.2) and the observation that assumptions 2.1 and 2.2 of this paper are satisfied in the present case.

For the cointegrated case, we make the following assumption.

Assumption 3.3 $|A(z)| = a_{st}(z)(1-z)^s$ *for* $0 < s \leq K$ *where* $a_{st}(z) \neq 0$ *for* $|z| \leq 1$. *The number* $r = K - s$ *is called the cointegrating rank of the series.*

Then the corresponding error-correction representation

$$y_t = \Pi y_{t-1} + \sum_{i=1}^{p_0-1} \Gamma_j \Delta y_{t-j} + m(L)u_t \tag{11}$$

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).

Specification

Since a legitimate critique of VARMA modeling is the increased specification uncertainty, we think that a serious forecast comparison has to involve modeling uncertainty. Therefore, we chose to select the orders data-dependent in our forecast study as described in the following.

First, the sample mean is subtracted from the observations as justified above. In order to determine the lag orders p and q of the VARMA model (1) with the FMA structure in (3) we apply a two-step approach similar to Dufour & Pelletier (2008) and use the information criterion they have suggested. Our procedure works as follows.

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

$$y_t = \sum_{i=1}^{h_T} \Pi_i^{h_T} y_{t-i} + u_t^{h_T}. \tag{12}$$

   Denote the estimated residuals from (12) by $\hat{u}_t^{h_T}$.

2. Regress y_t on $\phi_{t-1}^{h_T}(p,q) = [y'_{t-1}, \ldots, y'_{t-p}, \hat{u}_{t-1}^{h_T\prime}, \ldots, \hat{u}_{t-q}^{h_T\prime}]'$, $t = s_T + 1, \ldots T$, imposing the FMA restriction in (3) for all combinations of $p = k + 1 \leq p_T$ and $q \leq q_T$ with $s_T = \max(p_T, q_T) + h_T$ using OLS. Denote the estimate of the corresponding covariance error 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. Compute the information criterion

$$DP(p,q) = \ln |\hat{\Sigma}_T(p,q)| + \dim(\gamma^{(p,q)}) \frac{(\ln N)^{1+\nu}}{N}, \ \nu > 0 \tag{13}$$

   where $N = T - s_T$ and $\dim(\gamma^{(p,q)})$ is the dimension of the vector of free parameters of the corresponding VARMA(p,q) model in levels.

3. Choose the AR and MA orders by $\widehat{(p,q)}_{IC} = \operatorname{argmin}_{(p,q)} DP(p,q)$, where the minimization is over $\{1, 2, \ldots, p_T\} \times \{0, 1, \ldots, q_T\}$.

In order to show consistency we make the following assumption which is equivalent to Assumption A.2 in Poskitt (2003).

Assumption 3.4 *The true error term vectors* $u_t = (u'_{t,1}, u'_{t,2}, \ldots, u'_{t,K}), t = 1-p, \ldots, 0, 1, \ldots, T$*, 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.*

Using Assumption 3.4 we obtain the following theorem on the consistency of the order estimators.

Theorem 3.3 *If Assumptions 3.1-3.4 hold, if* $h_T = [c(\ln T)^a]$ *is the integer part of* $c(\ln T)^a$ *for some* $c > 0$*,* $a > 1$*, and if* $\max(p_T, q_T) < h_T$*, then the orders chosen according to* (13) *converge a.s. to their true values.*

Theorem 3.3 is the counterpart to Dufour & Pelletier (2008, 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 use the same penalty term $C_T = (\ln N)^{1+\nu}$, $\nu > 0$, as in the stationary VARMA case. However, the assumptions on the error terms have to be strengthened. In particular, an *i.i.d.* assumption is needed in contrast to the strong mixing assumption employed by Dufour & Pelletier (2008). Existing formal results of Poskitt & Lütkepohl (1995) and Huang & Guo (1990) show that weakening Assumption 3.4, e.g. making an appropriate martingale difference sequence assumption on u_t, leads to too low convergence orders for the estimators obtained in steps 1. and 2.. As a consequence, the penalty term needs to be stronger, e.g. one may set it to $C'_T = h_T C_T$. Nevertheless, Poskitt (2003) argues that it is likely that the needed convergence orders can be obtained under weaker conditions than those stated in Assumption 3.4.

The practitioner has to chose values for ν, h_T, p_T, and q_T satisfying the conditions contained in Theorem 3.3. We set $\nu = 0.2$ following Dufour & Pelletier (2008). As pointed out by Poskitt (2003) and Lütkepohl (2005, Chapter 14) no clear guideline exists on how to select h_T for the non-stationary case. We adopt the rule $h_T = \max\left(\max(p_T, q_T) + 1, (\ln T)^{1.25}\right)$ from Poskitt (2003) with $p_T = q_T = 4$. Choosing larger values for p_T and q_T left the results virtually unchanged. Alternatively, one may use Akaike's information criterion (AIC) to determine h_T resulting in the estimator $\hat{h}_T^{AIC}$, say. While Poskitt (2003) conjectures that $\hat{h}_T^{AIC}$ satisfies the condition on h_T given in Theorem 3.3, the latter has only been proven for the BIC by Bauer & Wagner (2005, Corollary 1).

Estimation

Given the estimated orders and residuals of the long autoregression (12) we obtain Poskitt's (2003) initial estimator as follows.

The cointegrated VARMA model (2) can be conveniently written as

$$\Delta y_t = \Pi' y_{t-1} + [\mathbf{\Gamma}\ \mathbf{M}]\mathbf{Z}_{t-1} + u_t, \tag{14}$$

where $\mathbf{\Gamma} = \text{vec}[\Gamma_1, \ldots, \Gamma_k]$, $\mathbf{M} = \text{vec}[M_1, \ldots, M_q]$ and $\mathbf{Z}_{t-1} = [\Delta y'_{t-1}, \ldots, \Delta y'_{t-k}, u'_{t-1}, \ldots, u'_{t-q}]'$.

Let $\mathbf{Z}_t^{h_T}$ be the matrix obtained from $\mathbf{Z}_t$ by replacing the u_t by $\hat{u}_t^{h_T}$. Let γ_1 be the vector of free parameters in vec$[\mathbf{\Gamma}\ \mathbf{M}]$ and the augmented vector $\gamma_2 = (\text{vec}(\Pi)',\ \gamma_1')'$. Identification restrictions are imposed by defining suitable matrices R, R_2 such that vec$([\mathbf{\Gamma}\ \mathbf{M}]) = R\gamma_1$ and vec$([\Pi\ \mathbf{\Gamma}\ \mathbf{M}]) = R_2\gamma_2$, respectively. Equipped with these definitions, one can write

$$\begin{aligned} \Delta y_t &= \left(y'_{t-1} \otimes I_K,\ \mathbf{Z}_{t-1}^{h_T\,\prime}\right) \text{vec}([\Pi\ \mathbf{\Gamma}\ \mathbf{M}]) + u_t \\ &= \left(y'_{t-1} \otimes I_K,\ \mathbf{Z}_{t-1}^{h_T\,\prime}\right) R_2\gamma_2 + u_t \\ &= X_t\gamma_2 + u_t. \end{aligned} \tag{15}$$

Poskitt's (2003) initial estimator is the feasible GLS estimator

$$\hat{\gamma}_2 = \left(\sum_{h_T+1}^{T} X_t'(\hat{\Sigma}_{\Delta,T})^{-1}X_t\right)^{-1} \sum_{h_T+1}^{T} X_t'(\hat{\Sigma}_{\Delta,T})^{-1}\Delta y_t, \tag{16}$$

which is strongly consistent (Poskitt 2003, Propositions 4.1 and 4.2) given Assumptions 3.1-3.4 and $\hat{\Sigma}_{\Delta,T}$ is an estimate of Σ_u obtained from OLS estimation of (15).[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

$$\hat{\alpha} = \hat{\Pi}[., 1:r], \tag{17}$$

$$\begin{aligned} \hat{\beta}^* &= \left(\hat{\alpha}'\left(\hat{M}(1)\hat{\Sigma}_T\hat{M}(1)'\right)^{-1}\hat{\alpha}\right)^{-1} \\ &\times \left(\hat{\alpha}'\left(\hat{M}(1)\hat{\Sigma}_T\hat{M}(1)'\right)^{-1}\hat{\Pi}[., r+1:K]\right). \end{aligned} \tag{18}$$

These estimates are taken as starting values for one iteration of a conditional maximum likelihood estimation procedure as in Yap & Reinsel (1995). Define the vector of free parameters, given the cointegration restrictions, as $\delta := (\text{vec}((\beta^*)')', \text{vec}(\alpha)', \gamma_1')'$ and its value at the jth iteration as $\delta^{(j)}$. The elements of the initial vector $\delta^{(0)} = \hat{\delta}$ correspond to (16) - (18). Compute $u_t^{(j)}$ and $\Sigma_u^{(j)}$ according to

$$M^{(j)}(L)u_t^{(j)} = \Delta y_t - \alpha^{(j)}(\beta^{(j)})'y_{t-1} - \Gamma^{(j)}(L)\Delta y_{t-1}, \tag{19}$$

$$\Sigma_u^{(j)} = \frac{1}{T}\sum_t^T u_t^{(j)}(u_t^{(j)})' \tag{20}$$

[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 $W_t^{(j)} := -\frac{\partial u_t^{(j)}}{\partial \delta_t'}$ is needed for computing one iteration of the proposed Newton-Raphson iteration. Also $W_t^{(j)}$ can be calculated iteratively as

$$(W_t^{(j)})' = \Big[(y_{t-1}' H \otimes \alpha),\ (y_{t-1}'\beta \otimes I_K),\ ((\mathbf{Z}_{t-1}^{(j)})' \otimes I_K)R\Big] - \sum_{i=1}^{q} M_i (W_{t-i}^{(j)})'$$

where $H' := [0_{((K-r)\times r)},\ I_{K-r}]$. The estimate is then updated according to

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

which amounts to a GLS estimation step. The estimates of the residuals and their covariance can be updated according to (19) and (20). The one-step iteration estimator $\delta^{(1)}$ is consistent and fully efficient asymptotically according to Yap & Reinsel (1995, Theorem 2) given the strong consistency of the initial estimator $\hat{\gamma}_2$ in (16).

Given estimates of the parameters and innovations, forecasts are obtained by using the implied VARMA form in levels. Finally, the sample mean, which was subtracted earlier, is added to the forecasts.

3.2 Benchmark Models

There are two benchmark models in the forecasting exercise. The first is the multivariate random walk $y_t = y_{t-1} + u_t, \quad u_t \sim i.i.d.(0_K, \Sigma_u)$, where the notation means that u_t is an independent white noise process. Point forecasts are obtained in a standard way.

The second benchmark model is the VECM

$$\Delta y_t = \mu_0 + \Pi y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + u_t, \quad t = k+2, \ldots, T, \tag{21}$$

with the assumptions on the initial values, parameters and the u_t are analogous to the assumptions made for the VARMA (1). The lag length is chosen by using the Bayesian information criterion, $\hat{p}_{BIC} = \text{argmin}_p\, BIC(p)$, where the minimization is over $p = 1, \ldots, p_T$ and $\hat{k}_{BIC} = \hat{p}_{BIC} - 1$. From the results in Bauer & Wagner (2005), we take $p_T = [(T/\log T)^{1/2}]$. Paulsen (1984) shows that the standard order selection criteria are consistent for multivariate autoregressive processes with unit roots. The BIC is

$$BIC(p) = \ln|\hat{\Sigma}(p)| + \ln N \frac{(p\,K+1)K}{N}, \tag{22}$$

where $N = T - p_T$, $\hat{\Sigma}(p) = \sum_{t=p_T+1}^{T} \hat{u}_t \hat{u}_t'/N$ is an estimate of the error term covariance matrix Σ and the $\hat{u}_t$ are obtained by estimating an unrestricted

VAR model of order p using $y_{p_T+1-p}, \ldots, y_T$ by OLS. After that, the parameters of (21) are estimated by reduced rank maximum likelihood estimation (Johansen 1988, 1991, 1996). Forecasts are obtained iteratively by using the implied estimated VAR form.

4 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. In order to show its potential usefulness, we applied the procedure in a forecasting exercise for US interest rates and found promising results.

There are a couple of issues which could be followed up. For example, the intercept term in the cointegrating relation is treated by subtracting the sample mean from the series and it would be desirable to have a more efficient method in this case. 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. & Vahid, F. (2008), 'VARMA versus VAR for macroeconomic forecasting', *Journal of Business & Economic Statistics* **26**, 237–252.

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

Campbell, J. Y. & Shiller, R. J. (1987), 'Cointegration and tests of present value models', *Journal of Political Economy* **95**(5), 1062–88.

Carstensen, K. (2003), 'Nonstationary term premia and cointegration of the term structure', *Economics Letters* **80**(3), 409–413.

Cavaliere, G., Rahbek, A. & Taylor, A. (2010), 'Co-integration rank testing under conditional heteroskedasticity', *Econometric Theory* **26**(6), 1719–1760.

Clarida, R. H., Sarno, L., Taylor, M. P. & Valente, G. (2006), 'The role of asymmetries and regime shifts in the term structure of interest rates', *Journal of Business* **79**(3), 1193–1224.

Cooley, T. F. & Dwyer, M. (1998), 'Business cycle analysis without much theory. A look at structural VARs', **83**, 57–88.

Cuthbertson, K. (1996), 'The expectations hypothesis of the term structure: The uk interbank market', *The Economic Journal* **106**(436), 578–592.

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

Dufour, J. M. & Pelletier, D. (2008), 'Practical methods for modelling weak VARMA processes: Identification, estimation and specification with a macroeconomic application', *Discussion Paper, McGill University, CIREQ and CIRANO* .

Engsted, T. & Tanggaard, C. (1994), 'Cointegration and the us term structure', *Journal of Banking & Finance* **18**(1), 167–181.

Fernández-Villaverde, J., Rubio-Ramírez, J. F., Sargent, T. J. & Watson, M. W. (2007), 'A,B,C's (and D)'s of understanding VARs', *American Economic Review* **97**(3), 1021–1026.

Feunou, B. (2009), 'A no-arbitrage VARMA term structure model with macroeconomic variables', *Working Paper, Duke University* . June 2009, Mimeo, Duke University.

Guo, L., Chen, H.-F. & Zhang, J.-F. (1989), 'Consistent order estimation for linear stochastic feedback control systems (carma model)', *Automatica* **25**(1), 147–151.

Hall, A. D., Anderson, H. M. & Granger, C. W. J. (1992), 'A cointegration analysis of treasury bill yields', *The Review of Economics and Statistics* **74**(1), 116–26.

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

Huang, D. & Guo, L. (1990), 'Estimation of nonstationary ARMAX models based on the Hannan-Rissanen method', *The Annals of Statistics* **18**(4), 1729–1756.

Johansen, S. (1988), 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control* **12**(2-3), 231–254.

Johansen, S. (1991), 'Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models', *Econometrica* **59**(6), 1551–1580.

Johansen, S. (1996), *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford.

Kascha, C. & Mertens, K. (2009), 'Business cycle analysis and VARMA models', *Journal of Economic Dynamics and Control* **33**(2), 267–282.

Lai, T. L. & Wei, C. Z. (1982), 'Asymptotic properties of projections with applications to stochastic regression problems', *Journal of Multivariate Analysis* **12**, 346–370.

Lütkepohl, H. (1984*a*), 'Linear aggregation of vector autoregressive moving average processes', *Economics Letters* **14**(4), 345–350.

Lütkepohl, H. (1984*b*), 'Linear transformations of vector ARMA processes', *Journal of Econometrics* **26**(3), 283–293.

Lütkepohl, H. (1996), *Handbook of Matrices*, Chichester: Wiley.

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

Lütkepohl, H. & Claessen, H. (1997), 'Analysis of cointegrated VARMA processes', *Journal of Econometrics* **80**(2), 223–239.

Monfort, A. & Pegoraro, F. (2007), 'Switching varma term structure models', *Journal of Financial Econometrics* **5**(1), 105–153.

Nielsen, B. (2006), Order determination in general vector autoregressions, *in* H.-C. Ho, C.-K. Ing & T. L. Lai, eds, 'Time Series and Related Topics: In Memory of Ching-Zong Wei', Vol. 52 of *IMS Lecture Notes and Monograph Series*, Institute of Mathematical Statistics, pp. 93–112.

Paulsen, J. (1984), 'Order determination of multivariate autoregressive time series with unit roots', *Journal of Time Series Analysis* **5**(2), 115–127.

Poskitt, D. (2006), 'On the identification and estimation of nonstationary and cointegrated ARMAX systems', *Econometric Theory* **22**(6), 1138–1175.

Poskitt, D. S. (2003), 'On the specification of cointegrated autoregressive moving-average forecasting systems', *International Journal of Forecasting* **19**(3), 503–519.

Poskitt, D. S. (2009), Vector autoregressive moving-average identification for macroeconomic modeling: Algorithms and theory, Working Paper 12/2009, Monash University.

Poskitt, D. S. & Lütkepohl, H. (1995), Consistent specification of cointegrated autoregressive moving average systems, Discussion Paper 1996-74, Humboldt Universität zu Berlin, SFB 373.

Shea, G. S. (1992), 'Benchmarking the expectations hypothesis of the interest-rate term structure: An analysis of cointegration vectors', *Journal of Business & Economic Statistics* **10**(3), 347–66.

Tiao, G. C. & Tsay, R. S. (1989), 'Model specificationi in multivariate time series', *Journal of the Royal Statistical Society, B* **51**(2), 157–213.

Yap, S. F. & Reinsel, G. C. (1995), 'Estimation and testing for unit roots in a partially nonstationary vector autoregressive moving average model', *Journal of the American Statistical Association* **90**(429), 253–267.

Appendix

Proof of Theorem 3.1:

$\Rightarrow$: Suppose $(y_t)_{t=1-m}^T$ satisfies (5) given initial conditions. One can view the sequence $(u_t)_{t=1-m}^T$ as a solution to (5) 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 $A_0 = -M_0\Pi_0$ and

$$A_i = \sum_{j=0}^{i} M_j \Pi_{i-j}, \text{ for } i = 1, \ldots, m \tag{23}$$

$$0 = \sum_{j=0}^{p} M_j \Pi_{i-j}, \text{ for } i \geq m+1 \tag{24}$$

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 $t \geq 1$

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

Further, define $(n_t)_{t=1-m}^T$ by $n_t = u_t^P - u_t$ for $t = 1-m, \ldots, 0$ and $0 = \sum_{i=0}^{m} M_j n_{t-i}$, for $t = 1, \ldots, T$.

By construction of n_t, $y_t = \sum_{s=1}^{t+m-1} \Pi_s y_{t-s} + u_t + n_t$ for $t = 1-m, \ldots, 0$. Then, we also have $u_t = u_t^P - n_t$ for $t \geq 1$ as $(-n_t)_{t=1-m}^T$ represents a solution to the homogeneous system.

$\Leftarrow$: Conversely, suppose $(y_t)_{t=1-m}^T$ admits an autoregressive representation as in (6) 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 $i \geq m+1$ 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

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

Which leads to

$$\begin{aligned} -\sum_{v=0}^{t+m-1}\sum_{j=0}^{\min(v,m)} M_j\Pi_{v-j}y_{t-v} &= -\sum_{v=0}^{m}\sum_{j=0}^{v} M_j\Pi_{v-j}y_{t-v} \\ &= A_0y_t - \sum_{v=1}^{m} A_v y_{t-v} = M(L)u_t \end{aligned}$$

where the last line *defines* the A_v's.

Proof of Theorem 3.2:

From Theorem 3.1, $(y_t)_{t=1-m}^{T}$ has a autoregressive representation. One can express conditions (9) and (10) for a suitable pair $[A(z),\, m(z)I_K]$ by defining the polynomials $\Pi(z) = -(\Pi_0 + \Pi_1 z + \ldots)$ and $n(z) = n_{1-m} + n_{-p}z + \ldots$. One also defines the (stochastic) polynomial $o(z) = o_0 + o_1 z + \ldots + o_{t_0}z^{\underline{t}}$ which captures that $\sum_{j=0}^{\max(q,t-m+1)} m_j n_{t-j} \neq 0$ for $t \leq \underline{t}$ for some $\underline{t}$.

$$A(z) = m(z)\Pi(z) \tag{25}$$

$$o(z) = m(z)n(z) \tag{26}$$

Then, for given polynomials $(\Pi(z), n(z))$ one defines the set of all scalar polynomials $m(z)$ with the first coefficient normalized to one for which there exist finite polynomials $A(z)$ and $o(z)$ such that the (25) and (26) are satisfied. Denote this set by S. Since the (normalized) determinant of $M(z)$ satisfies the above conditions, S is not empty. Denote one solution to

$$\min_{m(L)\in S} \deg(m(L)),$$

by $m_0(z)$ with degree q_0, where $\deg : S \to \mathbb{N}$ is the function that assigns the degree to every polynomial in S. Denote the associated polynomials by $A_0(z), o_0(z)$ with degrees p_0, t_0, respectively.

Suppose, there is another solution of the same degree $m_1(z) = 1 + m_{1,1}z + \ldots + m_{1,q_0}z^{q_0}$ with

$$\begin{aligned} A_1(z) &= m_1(z)\Pi(z) \\ o_1(z) &= m_1(z)n(z) \end{aligned}$$

Since both polynomials are of degree q_0, $a = m_{0,q_0}/m_{1,q_0}$ exists and one gets

$$\begin{aligned} (A_0(z) - aA_1(z)) &= (m_0(z) - am_1(z))\Pi(z) \\ (o_0(z) - ao_1(z)) &= (m_0(z) - am_1(z))n(z) \end{aligned}$$

Then, normalization of the first non-zero coefficient of $(m_0(z) - am_1(z))$ would give a polynomial in S with degree smaller than q_0, a contradiction. Thus $m_0(z)$ is unique.

Then, condition (25) alone would imply left-coprimeness of $[A_0(z),\, m_0(z)I_K]$ but if $n(z) \neq 0$ the minimal orders p_0, q_0 might well be above those of the left-coprime solution to (25).

Proof of Theorem 3.3:

Similar to Guo, Chen & 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 & 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 (13). Then

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

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:

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

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.

Then, one can write

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

in order to summarize the model in matrix notation by

$$\begin{aligned} Y_T &= X_T^0(p,q)\gamma(p,q) + U_T \\ &= X_T(p,q)\gamma(p,q) + [X_T^0(p,q) - X_T(p,q)]\gamma(p,q) + U_T \\ &= X_T(p,q)\gamma(p,q) + R_T + U_T, \end{aligned} \tag{27}$$

where

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

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.

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

The estimator of the error covariance matrix Σ 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

$$\begin{aligned} &\| \hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q) \| = \lambda_{\max}\left(\hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q)\right) \\ &\leq \text{tr}\left(\hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q)\right) = \text{tr}\left(\hat{\Sigma}_T(p_0,q_0)\right) - \text{tr}\left(\hat{\Sigma}_T(p,q)\right) \\ &= T^{-1}Z_T(p_0,q_0)'Z_T(p_0,q_0) - T^{-1}Z_T(p,q)'Z_T(p,q) \\ &= T^{-1}[R_T + U_T]'X_T(p,q)\left[X_T(p,q)'X_T(p,q)\right]^{-1} X_T(p,q)'[R_T + U_T] \\ &\quad - T^{-1}[R_T + U_T]'X_T(p_0,q_0)\left[X_T(p_0,q_0)'X_T(p_0,q_0)\right]^{-1} X_T(p_0,q_0)'[R_T + U_T] \end{aligned} \tag{28}$$

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

in (28)

$$\begin{aligned}
&[R_T + U_T]' X_T(p,q) \left[X_T(p,q)' X_T(p,q)\right]^{-1} X_T(p,q)'[R_T + U_T] \\
&\quad = O\left(|| \left[X_T(p,q)' X_T(p,q)\right]^{-1/2} X_T(p,q)'[R_T + U_T]||^2\right) \\
&\quad = O\left(|| \left[X_T(p,q)' X_T(p,q)\right]^{-1/2} X_T(p,q)' R_T||^2\right) \\
&\qquad + O\left(|| \left[X_T(p,q)' X_T(p,q)\right]^{-1/2} X_T(p,q)' U_T||^2\right),
\end{aligned} \tag{29}$$

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

As in Poskitt & Lütkepohl (1995, Proof of Theorem 3.2), we obtain from Lai & Wei (1982, Theorem 3) for any $m = max(p,q)$

$$\begin{aligned}
&|| \left[X_T(p,q)' X_T(p,q)\right]^{-1/2} X_T(p,q)' U_T||^2 \\
&\quad = O\left(\max\left\{1, \ln^+\left(\sum_{n=1}^{s}\sum_{t} ||y_{t-n}||^2 + ||\hat{u}_{t-n}^{h_T}||^2\right)\right\}\right) \\
&\quad = O(\ln\ m) + O\left(\ln\left(O\left\{\sum_t ||y_t||^2 + ||\hat{u}_t^{h_T}||^2\right\}\right)\right) a.s.,
\end{aligned} \tag{30}$$

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 & Lütkepohl (1995, Proof of Theorem 3.2, Proof of Lemma 3.1). Therefore, the second term on the r.h.s. of (30) is $O(\ln T)$ for all m. Hence, the left-hand side (l.h.s.) of (30) is $O(\ln T)$ $a.s.$ since $m \leq s_T \leq h_T = [c(\ln T)^a]$, $c > 0$, $a > 1$.

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

$$\begin{aligned}
|| \left[X_T(p,q)' X_T(p,q)\right]^{-1/2} X_T(p,q)' R_T||^2 &\leq \sum_{t=0}^{T-1}\sum_{i=1}^{K} r_{i,t}^2 \\
&\leq ||\gamma(p,q)||^2 \cdot \sum_{n=1}^{q}\sum_{t} ||u_{t-n} - \hat{u}_{t-n}^{h_T}||^2 \\
&= O(\ln T)\ a.s.,
\end{aligned} \tag{31}$$

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 $T^{-1}[R_T + U_T]' X_T(p,q) \left[X_T(p,q)' X_T(p,q)\right]^{-1} \times X_T(p,q)'[R_T + U_T] = O(\ln T/T)$ $a.s.$ uniformly in (p,q).

Using (29 - 31), we have $\| \hat{\Sigma}_T(p_0,q_0) - \hat{\Sigma}_T(p,q) \| = O(\ln T/T)$ 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 sufficently 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

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

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

$$\begin{aligned} \hat{\gamma}_T(p,q) &= \left[X_T'(p,q)X_T(p,q)\right]^{-1} X_T'(p,q)Y_T \\ &= [\text{vec}(\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_p)', \hat{m}_1, \hat{m}_2, \ldots, \hat{m}_q]'. \end{aligned} \tag{32}$$

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

$$\hat{\gamma}_T^0(p,q) = [\text{vec}(\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_{s_p})', \hat{m}_1, \hat{m}_2, \ldots, \hat{m}_{s_q}]' \tag{33}$$

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

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

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

$$\tilde{X}_T := (X_T^0(s_p,s_q) - X_T(s_p,s_q)), \tag{34}$$

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

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

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

$$\begin{aligned}
\hat{\Sigma}_T(p_0,q_0) &= \frac{1}{T}\sum_{t=1}^{T} z_t(p,q)z_t'(p,q) \\
&= \frac{1}{T}\sum_{t=1}^{T} x_t'(p_0,q_0)\tilde{\gamma}_T(p_0,q_0)\tilde{\gamma}_T'(p_0,q_0)x_t'(p_0,q_0) \\
&+ \frac{1}{T}\sum_{t=1}^{T} \left(x_t'(p_0,q_0)\tilde{\gamma}_T(p_0,q_0)\right)\left(u_t+\tilde{x}_t'\gamma(p_0,q_0)\right)' \\
&+ \frac{1}{T}\sum_{t=1}^{T} \left(u_t+\tilde{x}_t'\gamma(p_0,q_0)\right)\left(x_t'(p_0,q_0)\tilde{\gamma}_T(p_0,q_0)\right)' \\
&+ \frac{1}{T}\sum_{t=1}^{T} \left(u_t+\tilde{x}_t'\gamma(p_0,q_0)\right)\left(u_t+\tilde{x}_t'\gamma(p_0,q_0)\right)' \\
&= D_{1,T}^0 + (D_{2,T}^0 + (D_{2,T}^0)') + D_{3,T}^0,
\end{aligned}$$

where $D_{1,T}^0$, $D_{2,T}^0$, and $D_{3,T}^0$ are defined analogously. Then,

$$\begin{aligned}
\hat{\Sigma}_T(p,q) - \hat{\Sigma}_T(p_0,q_0) = D_{1,T} &+ \left(D_{2,T} + D_{2,T}' - D_{1,T}^0 - D_{2,T}^0 - (D_{2,T}^0)'\right) \\
&+ \left(D_{3,T} - D_{3,T}^0\right). \qquad (35)
\end{aligned}$$

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 (35) is o(1). We will further show that $D_{2,T}$, $D_{1,T}^0$, and $D_{2,T}^0$ are $o(1)$ *a.s.* and that $\liminf \lambda_{\max}(D_{1,T}) > 0$ while noting that $D_{1,T}$ is p.s.d. by construction, showing $\liminf \lambda_{\max}(\hat{\Sigma}_T(p,q) - \hat{\Sigma}_T(p_0,q_0)) > 0$, the desired result.

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

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

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

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

$$\begin{aligned}\tilde{\gamma}_T(p_0,q_0) &= \gamma(p_0,q_0) - \hat{\gamma}^0_T(p_0,q_0) \\ &= -\left[X_T'(p_0,q_0)X_T(p_0,q_0)\right]^{-1} X_T'(p_0,q_0)\left[\tilde{X}_T\gamma(p_0,q_0) + U_T\right]\end{aligned} \tag{36}$$

due to (27), (32), (33), and (34). Therefore,

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

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

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

where the last line follows from (29-31) of the first part of the proof; compare also Huang & Guo (1990, pp. 1754).

$D_{2,T}$**:** For $D_{2,T}$, we have

$$\begin{aligned}
&\| \frac{1}{T}\sum_{t=1}^{T}(x_t'(s_p,s_q)\tilde{\gamma}_T(p,q))(u_t+\tilde{x}_t'\gamma(s_p,s_q))' \| \\
&=\| \frac{1}{T}\sum_{t=1}^{T}\tilde{\Gamma}_T(p,q)\phi_t(s_p,s_q)(u_t+\tilde{x}_t'\gamma(s_p,s_q))' \| \\
&=\| \frac{1}{T}\tilde{\Gamma}_T(p,q)(\Phi_T'\Phi_T)^{1/2}(\Phi_T'\Phi_T)^{-1/2}\sum_{t=1}^{T}\phi_t(s_p,s_q)(u_t+\tilde{x}_t'\gamma(s_p,s_q))' \|,
\end{aligned}$$

where $\Phi_T := [\phi_0(s_p,s_q),\ldots,\phi_{T-1}(s_p,s_q)]'$ is a $T\times(s_p+s_q)\cdot K$ matrix. Then, similar to the approach in Huang & Guo (1990),

$$\begin{aligned}
&||\frac{1}{T}\sum_{t=1}^{T}(x_t'(s_p,s_q)\tilde{\gamma}_T(p,q))(u_t+\tilde{x}_t'\gamma(s_p,s_q))'|| \\
&\quad\leq \frac{1}{T}||\tilde{\Gamma}_T(p,q)(\Phi_T'\Phi_T)^{1/2}||\;||(\Phi_T'\Phi_T)^{-1/2}\Phi_T'(U^T+X^T)||,
\end{aligned}$$

where $U^T := [u_1,\ldots,u_T]'$, $X^T := [\tilde{x}_1'\gamma(s_p,s_q),\ldots,\tilde{x}_T'\gamma(s_p,s_q)]'$. Now, tedious but straightforward calculations lead to

$$\begin{aligned}
&\frac{1}{T}||\tilde{\Gamma}_T(p,q)(\Phi_T'\Phi_T)^{1/2}||\;||(\Phi_T'\Phi_T)^{-1/2}\Phi_T'(U^T+X^T)|| \\
&\quad\leq \left(\frac{1}{T}\tilde{\Gamma}_T(p,q)(\Phi_T'\Phi_T)\tilde{\Gamma}_T'(p,q)\right)^{1/2} \qquad (37)\\
&\qquad\times\left(\frac{1}{T}(U_T+\tilde{X}_T\gamma)'(\Phi_T\otimes I_K)((\Phi_T'\Phi_T)^{-1}\otimes I_K)(\Phi_T'\otimes I_K)(U_T+\tilde{X}_T\gamma)\right)^{1/2} \\
&\quad= [O(1)]^{1/2}\left[O\left(\frac{1}{T}\ln T\right)\right]^{1/2} = o(1)\ a.s.
\end{aligned}$$

following from the results on $D_{1,T}$ and again from (29-31) of the first part of the proof. Note in this respect that the results in (30) and (31) also hold when using the regressor matrix $\Phi_T\otimes I_K$ appearing in (37). 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 $D_{2,T}$, we can write

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

using arguments identical to those used to evaluate $D^0_{1,T}$ and noting that $T^{-1}(U_T+\tilde{X}_T'\gamma(p_0,q_0))'(U_T+\tilde{X}_T'\gamma(p_0,q_0))=T^{-1}U_T'U_T+o(1)=O(1)\ a.s.$ due to the results of Poskitt & Lütkepohl (1995, Proof of Theorem 3.2) and applying Poskitt (2003, Proposition 3.1). This completes the proof.

Table 1: Comparison of mean squared prediction errors

	RW				VECM				VARMA				VARMA YP			
	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12
3M	0.042	0.228	0.681	2.100	0.791	0.834	0.880	0.924	0.740	0.793	0.834	0.888	0.738	0.789	0.822	0.878
6M	0.042	0.235	0.676	2.015	0.832	0.954	1.001	1.004	0.794	0.903	0.944	0.966	0.780	0.886	0.926	0.954
3M	0.042	0.228	0.681	2.100	0.860	0.912	0.873	0.877	0.725	0.746	0.774	0.808	0.728	0.736	0.751	0.783
1Y	0.055	0.285	0.756	2.067	0.899	1.089	1.058	1.051	0.793	0.934	0.978	0.991	0.786	0.908	0.949	0.964
3M	0.042	0.228	0.681	2.100	1.062	1.065	0.961	0.896	0.817	0.886	0.899	0.847	0.803	0.908	0.912	0.825
5Y	0.065	0.282	0.591	1.120	0.930	1.076	1.042	1.078	0.879	0.987	1.030	1.063	0.875	0.979	1.016	1.032
3M	0.042	0.228	0.681	2.100	1.092	1.070	0.988	0.894	0.843	0.955	0.962	0.867	0.847	1.003	1.001	0.861
10Y	0.053	0.206	0.420	0.755	0.939	1.069	1.053	1.053	0.915	1.015	1.044	1.052	0.914	1.006	1.027	1.020
3M	0.042	0.228	0.681	2.100	0.779	0.800	0.889	0.934	0.795	0.821	0.877	0.902	0.794	0.829	0.878	0.899
6M	0.042	0.235	0.676	2.015	0.769	0.897	1.000	1.009	0.818	0.900	0.971	0.972	0.823	0.911	0.973	0.969
1Y	0.055	0.285	0.756	2.067	0.846	1.027	1.131	1.126	0.884	0.997	1.076	1.075	0.887	1.004	1.074	1.071
1Y	0.055	0.285	0.756	2.067	1.070	1.347	1.306	1.031	0.962	1.155	1.112	0.994	0.972	1.175	1.141	1.010
5Y	0.065	0.282	0.591	1.120	0.954	1.046	1.020	0.961	0.917	0.996	0.959	0.974	0.935	1.006	0.977	0.991
10Y	0.053	0.206	0.420	0.755	0.955	1.006	0.959	0.926	0.924	0.983	0.952	0.973	0.940	0.990	0.968	0.993
3M	0.042	0.228	0.681	2.100	0.911	1.088	1.173	1.013	0.728	0.783	0.829	0.838	0.730	0.802	0.856	0.838
1Y	0.055	0.285	0.756	2.067	0.979	1.297	1.372	1.167	0.822	1.004	1.042	0.999	0.830	1.029	1.070	0.996
10Y	0.053	0.206	0.420	0.755	0.960	1.073	1.093	1.085	0.912	1.006	1.025	1.048	0.911	1.002	1.013	1.018
3M	0.042	0.228	0.681	2.100	0.998	1.136	1.089	0.884	1.211	1.154	1.007	0.855	1.242	1.181	1.021	0.853
6M	0.042	0.235	0.676	2.015	1.027	1.186	1.162	0.951	1.193	1.161	1.057	0.913	1.222	1.185	1.071	0.911
1Y	0.055	0.285	0.756	2.067	1.072	1.217	1.194	1.010	1.195	1.164	1.079	0.967	1.216	1.183	1.091	0.966
5Y	0.065	0.282	0.591	1.120	1.023	1.045	1.019	0.988	1.059	1.012	0.959	0.960	1.065	1.016	0.961	0.957
10Y	0.053	0.206	0.420	0.755	1.011	1.014	0.982	0.960	1.017	0.986	0.948	0.939	1.017	0.984	0.943	0.933

Note: The table reports mean squared prediction errors for systems with different maturities and different models. VARMA refers to the VARMA model estimated via Poskitt's (2003) method and VARMA YP refers to the same model estimated via the method of Yap & Reinsel (1995).

Table 2: Comparison of determinant of mean squared prediction error matrices

	RW				VECM				VARMA				VARMA YP			
	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12
3M,6M	0.000	0.004	0.015	0.079	0.805	0.788	0.757	0.559	0.746	0.767	0.755	0.549	0.755	0.816	0.793	0.560
3M,1Y	0.001	0.013	0.067	0.344	0.857	0.860	0.707	0.551	0.688	0.734	0.657	0.495	0.705	0.774	0.666	0.484
3M,5Y	0.002	0.041	0.238	1.129	1.000	1.107	0.958	0.803	0.740	0.895	0.905	0.759	0.725	0.929	0.926	0.740
3M,10Y	0.002	0.037	0.213	1.117	1.032	1.172	1.067	0.860	0.794	1.026	1.046	0.840	0.792	1.073	1.076	0.817
3M,6M,1Y	0.000	0.000	0.000	0.002	0.750	0.737	0.801	0.603	0.751	0.686	0.745	0.562	0.758	0.694	0.748	0.561
3M,1Y,10Y	0.000	0.000	0.002	0.011	1.118	1.359	1.171	0.816	0.931	1.090	1.030	0.847	0.958	1.133	1.065	0.857
1Y,5Y,10Y	0.000	0.001	0.006	0.055	0.939	1.158	0.995	0.490	0.749	0.913	0.817	0.461	0.776	1.009	0.857	0.437
3M-10Y	0.000	0.000	0.000	0.000	0.927	1.051	0.943	0.449	1.147	1.057	0.928	0.491	1.170	1.069	0.923	0.470

Note: The table reports the determinant of the mean squared prediction error matrices for systems with different maturities and different models. VARMA refers to the VARMA model estimated via Poskitt's (2003) method and VARMA YP refers to the same model estimated via the method of Yap & Reinsel (1995).

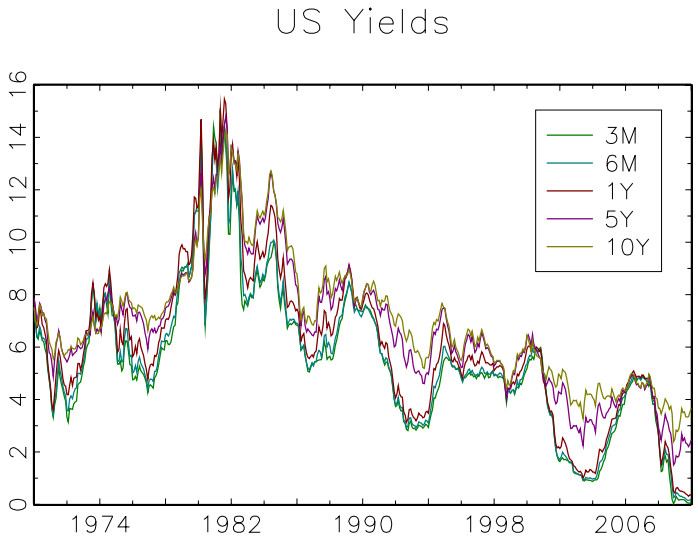
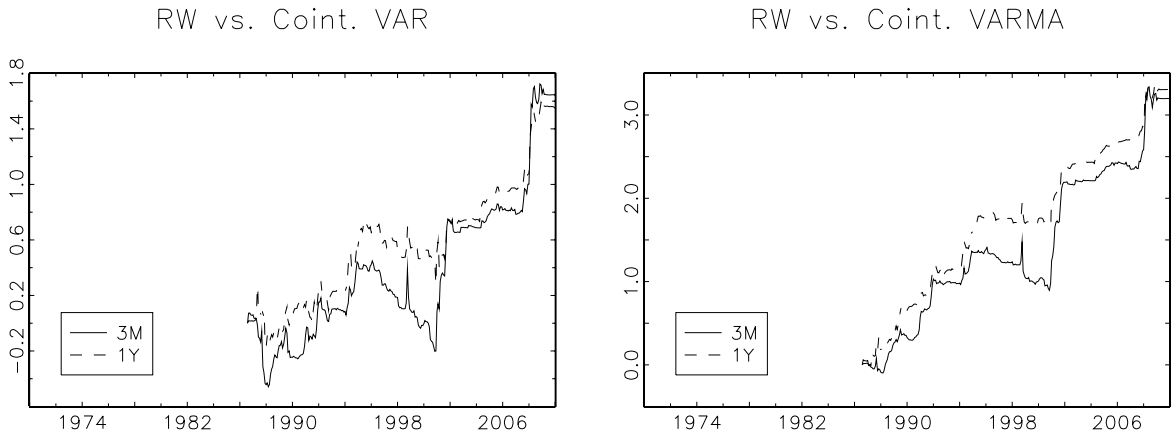
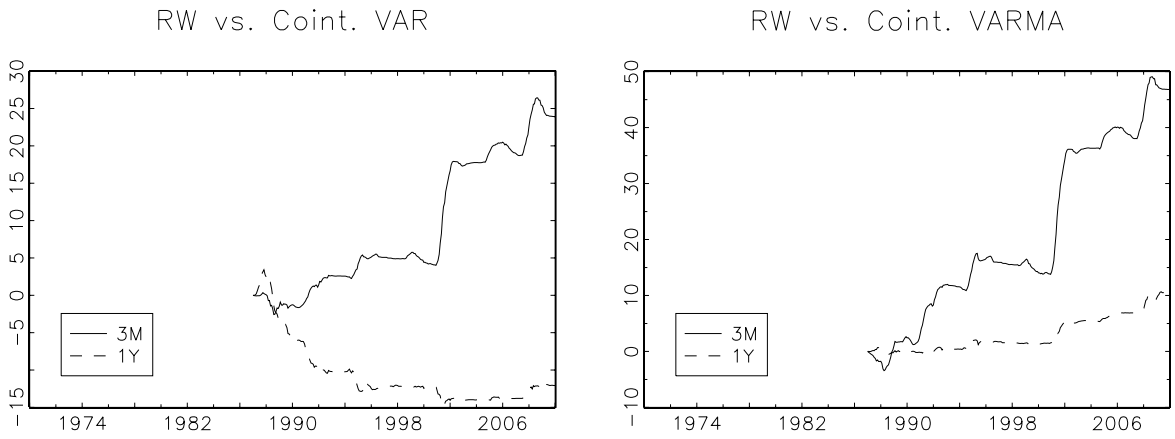


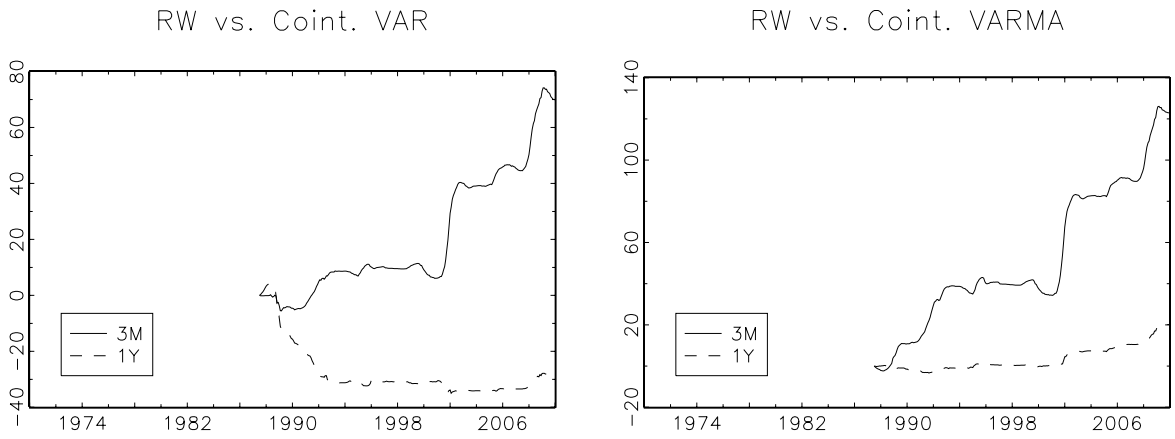
Figure 1: US treasury bills and bonds yields. See text for definitions.



(a) Forecasting horizon: 1 month



(b) Forecasting horizon: 6 month



(c) Forecasting horizon: 12 month

Figure 2: Cumulative squared prediction errors of the VECM and VARMA model for different horizons.