AARHUS UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Estimation and Forecasting in Vector Autoregressive Moving Average Models for Rich Datasets

Gustavo Fruet Dias and George Kapetanios

CREATES Research Paper 2014-37

Department of Economics and Business
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Estimation and Forecasting in Vector Autoregressive Moving Average Models for Rich Datasets

Gustavo Fruet Dias*
Department of Economics and Business Economics, Aarhus University
CREATES

George Kapetanios
King's College London

Abstract

We address the issue of modelling and forecasting macroeconomic variables using rich datasets, by adopting the class of Vector Autoregressive Moving Average (VARMA) models. We overcome the estimation issue that arises with this class of models by implementing an iterative ordinary least squares (IOLS) estimator. We establish the consistency and asymptotic distribution of the estimator for strong and weak VARMA(p,q) models. Monte Carlo results show that IOLS is consistent and feasible for large systems, outperforming the MLE and other linear regression based efficient estimators under alternative scenarios. Our empirical application shows that VARMA models are feasible alternatives when forecasting with many predictors. We show that VARMA models outperform the AR(1), BVAR and factor models, considering different model dimensions.

JEL classification numbers: C13, C32, C53, C63, E0

Keywords: VARMA, weak VARMA, weak ARMA, Forecasting, Rich and Large datasets, Iterative ordinary least squares (IOLS) estimator, Asymptotic contraction mapping.

Acknowledgments: This paper previously circulated under the title "Forecasting Medium and Large Datasets with Vector Autoregressive Moving Average (VARMA) Models". We are indebted to Emmanuel Guerre and the seminar participants at Queen Mary (Econometrics Reading Group), the 14th Brazilian Time Series and Econometrics School (Gramado, Aug 2011), the Econometric Society Australasian Meeting (Adelaide, July 2011), the Royal Economic Society Annual Conference (Royal Holloway, University of London, April 2011) and the 4th International Conference on Computational and Financial Econometrics (London, Dec 2010).

Gustavo Fruet Dias acknowledges support from CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation. The usual disclaimer applies.

*Corresponding author at: Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark. E-mail: gdias@econ.au.dk.

1 Introduction

The use of large arrays of economic indicators to forecast key macroeconomic variables has become very popular recently. Economic agents consider a wide range of information when they construct their expectations about the behavior of macroeconomic variables such as interest rates, industrial production, and inflation. In the past several years, this information has become more widely available through a large number of indicators that aim to describe different sectors and fundamentals from the whole economy. To improve forecast accuracy, large sized datasets that attempt to replicate the set of information used by agents to make their decisions are incorporated into econometric models.

For the past twenty years, macroeconomic variables have been forecasted using vector autoregression (VAR) models. This type of model performs well when the number of variables in the system is relatively small. When the number of variables increases, however, the performance of VAR forecasts deteriorates very fast, generating the so-called “curse of dimensionality”. In this paper, we propose the use of vector autoregressive moving average (VARMA) models, estimated with the iterative ordinary least squares (IOLS) estimator, as a feasible method to address the “curse of dimensionality” on medium and large sized datasets and improve forecast accuracy of macroeconomic variables. As far as our knowledge goes, the VARMA methodology has never been applied to medium and large sized datasets, as we do in this paper.

VARMA models have been studied for the past thirty years, but they have not been, by far, as popular as VAR models because of estimation and specification issues. Despite having attractive theoretical properties, estimation of VARMA models remains a challenge. Linear estimators (Hannan and Rissanen (1982), Hannan and Kavalieris (1984), Dufour and Jouini (2014), among others) and Bayesian methods (Chan, Eisenstat, and Koop (2016)) have been proposed in the literature as a way to overcome the numerical difficulties posed by the efficient maximum-likelihood estimator. We address the estimation issue and hence contribute to this literature by proposing the use of the IOLS estimator that is feasible for high dimensional VARMA models.

Other methodologies have been proposed in the literature to deal with the “curse of dimensionality”. We can divide them mainly in two groups of models. The first aims to overcome the dimensionality issue by imposing restrictions on the parameter matrices of a standard VAR model. Among the many important contributions from this field, we point out the following classes of models: Bayesian VAR (BVAR) (De Mol, Giannone, and Reichlin (2008) and Bańbura, Giannone, and Reichlin (2010)); Ridge (De Mol, Giannone, and Reichlin (2008)); reduced rank VAR (Carriero, Kapetanios, and Marcellino (2011)); and Lasso (De Mol, Giannone, and Reichlin (2008) and Tibshirani (1996)). The second group of models dealing with the "curse of dimensionality" reduces the dimension of the dataset by constructing summary proxies from the large dataset. Chief among these models is the class of factor models. The seminal works in this area are Forni, Hallin, Lippi, and Reichlin (2000) Stock and Watson (2002a) and Stock and Watson (2002b). Common factor models improve forecast accuracy and produce theoretically well-behaved impulse response functions, as reported by De Mol, Giannone, and Reichlin (2008) and Bernanke, Boivin, and Eliasz (2005).

VARMA models are able to capture two important features from these two groups of models. The first is the reduction of the model dimensionality, achieved by setting some elements of the parameter matrices to zero following uniqueness requirements. This produces parameter matrices which are not full rank, resembling the reduced rank VAR model of Carriero, Kapetanios, and Marcellino (2011). The second is the parsimonious summarizing of high-order autoregressive lags into low-order lagged shocks. By adopting VARMA, we allow lagged shocks from most of the macroeconomic variables in our dataset to play important roles in forecasting the future realizations of key macroeconomic variables. Additionally, VARMA models are closed under linear transformation and marginalization (see Lütkepohl (2007, Section 11.6.1)), providing additional flexibility and potentially better forecast performance. Moreover, as Dufour and Stevanović (2013) show, if the latent common factors follow either a finite VAR or VARMA processes, then the observed series will have a VARMA representation. This fact reinforces the use of VARMA model as a suitable framework to forecast key macroeconomic variables using potentially many predictors.

With regard to the theory, we establish the consistency and asymptotic distribution of the IOLS estimator for the strong and weak univariate ARMA(1,1) and VARMA(p,q) models.[1] Our asymptotic results are obtained under mild assumptions using the asymptotic contraction mapping framework of Dominitz and Sherman (2005). Albeit not efficient, we show through an extensive Monte Carlo study that the IOLS estimator has a good finite sample performance when compared to alternative estimators, such as the efficient MLE,

[1]ARMA(p,q) models are said to be weak processes if the disturbances are neither independent and identically distributed (*iid*), nor martingale difference sequence (*mds*), but only an uncorrelated process.

and important linear competitors. We report results from eighty six different simulations, considering different system dimensions, sample sizes, strong and weak innovations, and eigenvalues associated with the parameter matrices. We find the IOLS estimator performs well in a variety of scenarios, such as: small sample size; the eigenvalues associated with the parameter matrices are near-to-zero; and high-dimensional systems.

In the empirical part of this paper, we focus on forecasting three key macroeconomic variables: industrial production, interest rate, and CPI inflation using potentially large sized datasets. As in Carriero, Kapetanios, and Marcellino (2011), we use the 52 US macroeconomic variables taken from the dataset provided by Stock and Watson (2006) to construct systems with five different dimensions: 3, 10, 20, 40 and 52 variables. By doing that, we are able to evaluate the tradeoff between forecast gains by incorporating more information (large sized datasets) and the estimation cost associated with it. Additionally, the different system dimensions play the role of robustness check. We show that VARMA models are strong competitors and produce more accurate forecasts than the benchmark models (AR(1), BVAR and factor models) in different occasions. This conclusion holds for different system sizes and horizons.

The paper is structured as follows. In Section 2, we discuss the properties and identification of VARMA models and derive the IOLS estimator. In Section 3, we establish the consistency and asymptotic distribution of the IOLS estimator considering general strong and weak VARMA(p,q) models. In Section 4, we address the consistency and efficiency of VARMA models estimated with the IOLS procedure through a Monte Carlo study. In Section 5, we display the results from our empirical application. The proofs, tables and graphs are relegated to an Appendix and an online Supplement. Specifically, the Appendix brings the proofs for the Theorems, selected tables from the Monte Carlo study, and a summary of the empirical results covering all system sizes. The online Supplement brings the proof of Corollary 1, the auxiliary Lemmas 3-8 and proofs, the entire set of tables with results from the Monte Carlo and the empirical studies, and further discussion on selected topics.

2 VARMA Models, Identification and Estimation Procedures

Our interest lies in modelling and forecasting key elements of the K dimensional vector process $Y_t = (y_{1,t}, y_{2,t}, ..., y_{K,t})'$, where K is allowed to be large. We assume, as a baseline model, a general nonstandard VARMA(p,q) model where the means have been removed,[2]

$$A_0 Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + ... + A_p Y_{t-p} + M_0 U_t + M_1 U_{t-1} + ... + M_q U_{t-q}, \tag{1}$$

$$Y = BX + U. \tag{2}$$

The disturbances $U_t = (u_{1,t}, u_{2,t}, ..., u_{K,t})'$ are assumed to be a zero-mean white-noise process with a non-singular covariance matrix Σ_u, Y has dimension $(K \times T)$; $B = [(I_K - A_0), A_1, ..., A_p, (M_0 - I_K), M_1, ..., M_q]$ concatenates the parameter matrices with dimension $(K \times K(p+q+2))$; $X = (X_1, ..., X_T)$ can be seen as the matrix of regressors with dimension $(K(p+q+2) \times T)$, where $X_t = [Y_t, Y_{t-1}, ..., Y_{t-p}, U_t, U_{t-1}, ..., U_{t-q}]'$; and U is a $(K \times T)$ matrix of disturbances. Our baseline model is assumed to be stable and invertible, and the latter is crucial in our estimation process.[3]

2.1 Identification and Uniqueness

Nonstandard VARMA models require restrictions on the parameter matrices to ensure that the model is uniquely identified. Define the lag polynomials $A(L) = A_0 - A_1 L - A_2 L^2 - ... - A_p L^p$ and $M(L) = M_0 + M_1 L + M_2 L^2 + ... + M_q L^q$, where L is the usual lag operator. More specifically, we say that a model is unique if there is only one pair of stable and invertible polynomials $A\,(L)$ and $M\,(L)$, respectively, which satisfies the canonical MA representation

$$Y_t = A\,(L)^{-1} M\,(L) = \Theta\,(L)\,U_t = \sum_{i=0}^{\infty} \Theta_i U_{t-i}, \tag{3}$$

for a given $\Theta\,(L)$ operator. In contrast to the reduced form VAR models, setting $A_0 = M_0 = I_K$ is not sufficient to ensure an unique VARMA representation. Uniqueness of (1) is guaranteed by imposing restrictions on the $A\,(L)$ and $M\,(L)$ operators. The first require-

[2] We adopt the terminology in Lütkepohl, 2007, pg. 448 and call a stable and invertible VARMA representation as (1) nonstandard when A_0 and M_0 are allowed to be nonidentity invertible matrices. If $A_0 = M_0 = I_K$, we call it a standard VARMA model.

[3] A general VARMA(p,q) is considered stable and invertible if $\det\left(A_0 - A_1 z - A_2 z^2 - ... - A_p z^p\right) \neq 0$ $for\ |z| \leq 1$ and $\det\left(M_0 - M_1 z - M_2 z^2 - ...\ - M_p z^q\right) \neq 0$ $for\ |z| \leq 1$ hold, respectively.

ment is that the operators $A(L)$ and $M(L)$ are left-coprime, which implies that there is no left common factor, except for unimodular operators, that satisfies $C(L)\left[\bar{A}(L) : \bar{M}(L)\right] = [A(L) : M(L)]$.[4] It is, however, still necessary to impose further restrictions on $C(L)$ so that the left-coprime operators $A(L)$ and $M(L)$ are unique and hence the nonstandard VARMA representation in (1). This is achieved by restricting $C(L) = I_K$. Therefore, imposing a set of restrictions so that the $A(L)$ and $M(L)$ operators are unique left-coprime operators suffices to ensure uniqueness of stable and invertible nonstandard VARMA models.

A number of different strategies can be implemented to obtain unique VARMA representations, such as the extended scalar component approach of Athanasopoulos and Vahid (2008), the final equations form (see Lütkepohl, 2007, pg.362) and the Echelon form transformation (see Hannan and Kavalieris (1984), Lütkepohl and Poskitt (1996), among others). Athanasopoulos, Poskitt, and Vahid (2012) show that VARMA models specified using scalar components perform slightly better in empirical exercises than the ones using the Echelon form methodology. The authors claim, however, that the latter has the advantage of having a simpler identification procedure. More specifically, the Echelon form identification strategy can be fully automated, which turns to be a great advantage when dealing with medium and large systems. Moreover, the Echelon form transformation has the advantage, compared to final equations, to provide a more parsimonious parametrization, which is a highly desired feature when modelling medium and large systems. In this paper, we implement the Echelon form transformation as a way to impose uniqueness in both Monte Carlo and empirical applications. Because the three identification strategies (scalar components, final equations form and Echelon form) impose uniqueness through a set of linear restrictions on the $A(L)$ and $M(L)$ operators, the IOLS estimator can be directly implemented no matter what identification strategy the researcher chooses.

A general VARMA model such as the one stated in (1) is considered to be in its Echelon form if the conditions stated in equations (4), (5), (6) and (7) are satisfied (see Lütkepohl,

[4] $C(L)$ is an unimodular operator if $|C(L)|$ is a nonzero constant that does not depend on L.

2007, pg. 452 and Lütkepohl and Poskitt (1996) for more details):

$$p_{ki} := \begin{cases} \min\left(p_k+1, p_i\right) \text{ for } k \geq i \\ \min\left(p_k, p_i\right) \text{ for } k < i, \end{cases} \qquad \text{for } k, i = 1, ..., K, \tag{4}$$

$$\alpha_{kk}\left(L\right) = 1 - \sum_{j=1}^{p_k} \alpha_{kk,j} L^j \qquad \text{for } k = 1, ..., K, \tag{5}$$

$$\alpha_{ki}\left(L\right) = - \sum_{j=p_k-p_{ki}+1}^{p_k} \alpha_{ki,j} L^j \qquad \text{for } k \neq i, \tag{6}$$

$$m_{ki}\left(L\right) = \sum_{j=0}^{p_k} m_{ki,j} L^j, \qquad \text{for } k, i = 1, ..., K \qquad \text{with} \quad M_0 = A_0, \tag{7}$$

where $A(L)=[\alpha_{ki}]_{k,i=1,...,K}$ and $M(L)=[m_{ki}]_{k,i=1,...,K}$ are, respectively, the operators from the autoregressive and moving average components of the VARMA process. The arguments p_k for $k = 1, ..., K$ are Kronecker indices and denote the maximum degrees of both polynomials $A(L)$ and $M(L)$, being exogenously defined. We define $\mathbf{p} = (p_1, p_2, ..., p_K)'$ as the vector collecting the Kronecker indices. The p_{ki} numbers can be interpreted as the free coefficients in each operator $\alpha_{ki}\left(L\right)$ for $i \neq k$ from the $A\left(L\right)$ polynomial. By imposing restrictions on the coefficient matrices due to the Echelon form transformation, VARMA models have the desirable feature that many of the coefficients from both the autoregressive and moving average matrices are equal to zero. The restrictions imposed by the Echelon form are necessary and sufficient for the unique identification of stable and invertible nonstandard VARMA models. This follows because the Kronecker index in the $k-$th row of $\left[A\left(L\right) : M\left(L\right)\right]$ only specifies the maximum degree of all operators, and hence further restrictions (potentially data dependent) could be added. However, throughout this entire study, we restrict ourselves to the necessary and sufficient restrictions imposed by the Echelon form. Furthermore, we always consider the Echelon forms which yield VARMA models that cannot be partitioned into smaller independent systems.[5]

Finally, it is important to note that the VARMA representation in (1) is not a structural VARMA (SVARMA) model in its classical definition (see Gourieroux and Monfort (2015)), because (1) is not necessarily driven by independent (or uncorrelated) shocks. To construct impulse response functions which depend on the structural shocks, additional identification

[5]See discussion and examples in the online Supplement (Section S. 3) on how the Echelon form restricts elements of the AR and MA parameter matrices in (1). See also Section S. 6.2 for the discussion on how Hannan-Kavalieris procedure yields estimates of the Kronecker indices.

restrictions to the ones required for uniqueness are necessary. In particular, as noted by Gourieroux and Monfort (2015), the structural shocks can usually be derived by imposing restrictions either on the contemporaneous correlation among the innovations in (1) (see the so-called B-model in Lütkepohl, 2007, pg. 362 and the general identification theory in Rubio-Ramrez, Waggoner, and Zha (2010)), or on the long-run impact matrix of the shocks (Blanchard and Quah (1989)) or by imposing sign restrictions on some impulse response functions (Uhlig (2005)). This identification issue is common to both VARMA and VAR models, and has been primarily explored in the context of VAR models.

Further, we note an additional identification issue that is particular to VARMA models. As discussed, in Gourieroux and Monfort (2015), structural VARMA models may suffer from the presence of non-invertible AR or MA matrix polynomials, potentially leading to nonfundamental VARMA representations. The causes and possible remedies of this identification issue are discussed in detail in Gourieroux and Monfort (2015). In this paper, however, we restrict ourselves to the stable and invertible VARMA models as in (1) and, further, do not attempt to identify the structural shocks driving the SVARMA models, as we feel that these issues are beyond the scope of the paper, since its focus is on analysing our new proposed estimator within the context of rich datasets. However, these issues are worthy of further investigation, within the context of IOLS, in future work.

2.2 Estimation

VARMA models, similar to their univariate (ARMA model) counterparts, are usually estimated using MLE procedures. Provided that the model in (1) is uniquely identified and disturbances U_t are normally distributed, MLE delivers consistent and efficient estimators. Although MLE seems to be very powerful at first glance, it presents serious problems when dealing with VARMA models that account for medium and large sized datasets. We overcome this issue by implementing an IOLS procedure in the spirit of Spliid (1983), Hannan and McDougall (1988) and Kapetanios (2003). Compared to the previous authors, this paper goes much further in three different directions: first, by establishing the asymptotic theory for the IOLS estimator for both ARMA(1,1) and VARMA(p,q) models under assumptions compatible with the quasi-maximum-likelihood (QMLE) estimator; second, by extending the consistency and asymptotic normality to the weak VARMA(p,q) case; third by showing, through an extensive Monte Carlo, that, under different scenarios, the IOLS estimator outperforms the MLE and other linear regression based estimators.

The IOLS framework consists of computing ordinary least squares (OLS) estimates of the parameters using estimates of the latent regressors. These regressors are computed recursively at each iteration using the OLS estimates. We assume that the model in (1) is expressed in its Echelon form and is therefore uniquely identified. Echelon form transformation implies that $A_0 = M_0$, which leads to a different specification of matrices in (8) when compared with the compact notation displayed in (2),

$$\operatorname{vec}(Y) = \left(X' \otimes I_K\right) \operatorname{vec}(B) + \operatorname{vec}(U). \tag{8}$$

We now have that $B = [(I_K - A_0), A_1, ..., A_p, M_1, ..., M_q]$ with dimension $(K \times K(p+q+1))$; $X = (X_{\bar{q}+1}, ..., X_T)$ is the matrix of regressors with dimension $(K(p+q+1) \times T - \bar{q})$, where $\bar{q} = \max\{p, q\}$ and $X_t = \operatorname{vec}(Y_t - U_t, Y_{t-1}, ..., Y_{t-p}, U_{t-1}, ..., U_{t-q})$; $Y = (Y_{\bar{q}+1}, ..., Y_T)$ has dimension $(K \times T - \bar{q})$; and $U = (U_{\bar{q}+1}, ..., U_T)$ is a $(K \times T - \bar{q})$ matrix of disturbances. Note that by setting the dimensions of the X and Y matrices as $(K(p+q+1) \times T - \bar{q})$ and $(K \times T - \bar{q})$, respectively, we explicitly highlight the fact that we lose $\bar{q}$ observations in finite sample. Also, recall from Section 2.1 that the matrices of parameters may not be full matrices because the Echelon form transformation can set many of their elements to zero. We obtain the free parameters in the model by rewriting the matrix of parameters $\operatorname{vec}(B)$ into the product of a $\left(K^2(p+q+1) \times n\right)$ deterministic matrix R and a $(n \times 1)$ vector β,

$$\operatorname{vec}(Y) = \left(X' \otimes I_K\right) R\beta + \operatorname{vec}(U), \tag{9}$$

where n denotes the number of free parameters in the model and β concatenates all the free parameters.

Using the invertibility condition, we can express a finite VARMA model as an infinite VAR, $Y_t = \sum_{i=1}^{\infty} \Pi_i Y_{t-i} + U_t$. We compute consistent estimates of U_t, denoted as $\widehat{U}_t^0$, by truncating the infinite VAR representation using some lag order, $\tilde{p}$, that minimizes some criterion. Following the results from Ng and Perron (1995)[6] and Dufour and Jouini (2005), choosing $\tilde{p}$ proportional to $\ln(T)$ delivers consistent estimates of U_t, so that $\widehat{U}_t^0 = Y_t -$

[6]Lemmas 4.1 and 4.2 in Ng and Perron (1995) show that determining the truncation lag proportional to $\ln(T)$ guarantees that the difference between the residuals from the truncated VAR model and the ones obtained from the infinite VAR process is $o_p\left(T^{-1/2}\right)$ uniformly in $\tilde{p}$.

$\sum_{i=1}^{\tilde{p}} \widehat{\Pi}_i Y_{t-i}$. Substitute $\widehat{U}_t^0$ into the matrix X in (9) and denote it $\widehat{X}^0$. Note that both $\widehat{X}^0$ and vec (Y) change their dimensions to $(K(p+q+1) \times T - \bar{q} - \tilde{p})$ and $(K(T - \bar{q} - \tilde{p}) \times 1)$, respectively. This happens exclusively on this first iteration, because $\tilde{p}$ observations are lost on the VAR($\tilde{p}$) approximation of U_t. The first iteration of the IOLS algorithm is obtained by computing the OLS estimator from the modified version of (9),

$$\widehat{\beta}^1 = \left[R' \left(\widehat{X}^0 \widehat{X}^{0\prime} \otimes I_K\right) R\right]^{-1} R' \left(\widehat{X}^0 \otimes I_K\right) \text{vec}(Y). \tag{10}$$

It is relevant to highlight that the first step of the IOLS algorithm is the two-stage Hannan-Rissanen (HR) algorithm formulated in Hannan and Rissanen (1982), and for which Dufour and Jouini (2005) show the consistency and asymptotic distribution.

We are now in a position to use $\widehat{\beta}^1$ to recover the parameter matrices $\widehat{A}_0^1, ..., \widehat{A}_p^1$, $\widehat{M}_1^1, ..., \widehat{M}_q^1$ and a new set of residuals $\widehat{U}^1 = \left(\widehat{U}_1^1, \widehat{U}_2^1, ..., \widehat{U}_T^1\right)$ by recursively applying

$$\widehat{U}_t^1 = Y_t - \left[\widehat{A}_0^1\right]^{-1} \left[\widehat{A}_1^1 Y_{t-1} - ... - \widehat{A}_p^1 Y_{t-p} - \widehat{M}_1^1 \widehat{U}_{t-1}^1 - ... - \widehat{M}_q^1 \widehat{U}_{t-q}^1\right], \text{ for } t = 1, ..., T, \tag{11}$$

where $Y_{t-\ell} = \widehat{U}_{t-\ell}^1 = 0$ for all $\ell \geq t$. Setting the initial values to zero when computing the residuals recursively on any j iteration is asymptotically negligible (see Lemma 4). Note that the superscript on the parameter matrices refers to the iteration in which those parameters are computed, and the subscript is the usual lag order. We compute the second iteration of the IOLS procedure by plugging $\widehat{U}_t^1$ into (9) yielding $\widehat{X}^1$. Note that $\widehat{X}^1 = (\widehat{X}_{\bar{q}+1}^1, ..., \widehat{X}_T^1)$, where $\widehat{X}_t^1 = [Y_t - \widehat{U}_t^1, Y_{t-1}, ..., Y_{t-p}, \widehat{U}_{t-1}^1, ..., \widehat{U}_{t-q}^1]'$, is a function of the estimates obtained in the first iteration: $\widehat{\beta}^1$. Similarly as in (10), we obtain $\widehat{\beta}^2$ and its correspondent set of residuals recursively as in (11). The j^{th} iteration of the IOLS estimator is thus given by

$$\widehat{\beta}^j = \left[R' \left(\widehat{X}^{j-1} \widehat{X}^{j-1\prime} \otimes I_K\right) R\right]^{-1} R' \left(\widehat{X}^{j-1} \otimes I_K\right) \text{vec}(Y). \tag{12}$$

We stop the IOLS algorithm when estimates of β converge. In both the empirical application and the Monte Carlo study, we assume that $\widehat{\beta}^j$ converges if $\| \widehat{U}^j - \widehat{U}^{j-1} \| \leq \epsilon$ holds from some exogenously defined criterion ϵ[7], where $\|.\|$ accounts for the Frobenius norm.

[7] The consistency and asymptotic normality results discussed in Section 3 require that $\frac{\ln(T)}{j} = o(1)$. By choosing ϵ sufficiently small, we also meet this theoretical requirement in both empirical application and Monte Carlo study.

In practise, we find that there are cases where the IOLS estimator does not converge. Section 3 discusses, using asymptotic arguments, the theoretical reasons why the IOLS estimator may not converge. From an empirical point of view, non-convergence of the IOLS estimator arises when some iteration of the algorithm generates either a non-invertible or a non-stable model. As general rule to identify whether the IOLS is going to converge, it suffices to check, on every iteration, the roots of the AR and MA polynomials. If all the roots from both AR and MA polynomials lie outside the unit circle, we continue with the algorithm as it is likely to converge. If the invertibility or stability conditions are violated at some iteration, we abort the algorithm and adopt the consistent HR estimator, $\widehat{\beta}^1$, given in (10). In both Monte Carlo simulations and empirical study we implement this general rule to define whether the IOLS estimator converges. We also notice from our simulations that sample size, number of free parameters, system dimension, and the true values of the free parameters play an important role on the convergence rates of the IOLS estimator. In general, we find that convergence rates increases monotonically with T, while large systems with a high number of free parameters are more likely to face convergence problems.

3 Theoretical Properties

This section provides theoretical results regarding the consistency and the asymptotic distribution of the IOLS estimator. A previous attempt to establish these results have been made by Hannan and McDougall (1988). They prove the consistency of the IOLS estimator considering the univariate ARMA(1,1) specification, but no formal result is provided for the asymptotic normality. Hence, our aim in this section is to formalize the consistency and the asymptotic normality of the IOLS estimator for the general strong and weak VARMA(p,q) models. For simplicity, we start our analysis on the univariate ARMA(1,1) model. As a second step, we extend the consistency and asymptotic normality results to the strong VARMA(p,q) model. Finally, we extend theoretical results to weak ARMA(1,1) and VARMA(p,q) models. Overall, this section differs from the work of Hannan and McDougall (1988) in important ways. First, we derive the asymptotic distribution for the general VARMA(p,q) model; second, we allow the errors to satisfy mixing conditions rather than imposing that they are independent and identically distributed, *iid*, or martingale difference sequence, *mds*, processes, yielding the consistency and asymptotic normality for the weak

ARMA and VARMA models; third, our theory explicitly accounts for the effects of setting initial values equal to zero when updating the residuals on each iteration; fourth, we show that the IOLS estimator converges globally to the sample fixed point, rather than locally as previously derived by Hannan and McDougall (1988); and fifth, we establish the necessary rates j has to increase, such that the consistency and the asymptotic normality results hold.

Similarly as in Section 2, we define our baseline VARMA(p,q) model expressed in its Echelon form as in (13) and its more compact notation in (14):

$$A_0 Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + ... + A_p Y_{t-p} + A_0 U_t + M_1 U_{t-1} + ... + M_q U_{t-q}, \tag{13}$$

$$Y_t = \left(X_t' \otimes I_K\right) R\beta + U_t. \tag{14}$$

We base our asymptotic results on the general theory for iterative estimators developed by Dominitz and Sherman (2005). Their approach relies on the concept of the Asymptotic Contraction Mapping (ACM). Denote $(\mathbb{B}, d)$ as a metric space where $\mathbb{B}$ is the closed ball centered in β and d is a distance function; $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space, where Ω is a sample space, $\mathcal{A}$ is a σ-field of subsets of Ω and $\mathbb{P}$ is the probability measure on $\mathcal{A}$; and $K_T^{\omega}(.)$ is a function defined on $\mathbb{B}$, with $\omega \in \Omega$. From the definition in Dominitz and Sherman, 2005, pg. 841, *"The collection $\{K_T^{\omega}(.) : T \geq 1, \omega \in \Omega\}$ is an ACM on $(\mathbb{B}, d)$ if there exist a constant $c \in [0, 1)$ that does not depend on T or ω, and sets $\{\mathcal{A}_T\}$ with each $\mathcal{A}_T \subseteq \Omega$ and $\mathbb{P}\mathcal{A}_T \to 1$ as $T \to \infty$, such that for each $\omega \in \mathcal{A}_T$, $K_T^{\omega}(.)$ maps $\mathbb{B}$ to itself and for all $x, y \in \mathbb{B}$, $d\left(K_T^{\omega}(x), K_T^{\omega}(y)\right) \leq cd(x, y)$"*. As pointed out by Dominitz and Sherman (2005), if a collection is an ACM, then it will have a unique fixed point in $(\mathbb{B}, d)$, where the fixed point now depends on the sample characteristics, i.e., of each T and ω. As discussed in Dominitz and Sherman, 2005, pg. 840, their ACM definition nests the case where the population mapping is a fixed deterministic function.

Definition 1 (General Mapping) *We define the sample mapping $\widehat{N}_T\left(\widehat{\beta}^j\right)$ and its population counterpart $N\left(\beta^j\right)$ as follows:*

i. $\widehat{\beta}^{j+1} = \widehat{N}_T\left(\widehat{\beta}^j\right) = \left[\frac{1}{T-\bar{q}} \sum_{t=\bar{q}+1}^{T} \widetilde{X}_t^{j\prime} \widetilde{X}_t^j\right]^{-1} \left[\frac{1}{T-\bar{q}} \sum_{t=\bar{q}+1}^{T} \widetilde{X}_t^{j\prime} Y_t\right]$,

ii. $\beta^{j+1} = N\left(\beta^j\right) = \mathbb{E}\left[\widetilde{X}_{\infty,t}^{j\prime} \widetilde{X}_{\infty,t}^j\right]^{-1} \mathbb{E}\left[\widetilde{X}_{\infty,t}^{j\prime} Y_t\right]$,

where $\widetilde{X}_t^j = \left[\left(\widehat{X}_t^{j\prime} \otimes I_K\right) R\right]$ and $\widetilde{X}_{\infty,t}^j = \left[\left(X_{\infty,t}^{j\prime} \otimes I_K\right) R\right]$ have dimensions $(K \times n)$ and denote the regressors computed on the j^{th} iteration; n is the number of free parameters in the model; $\widehat{X}_t^j = \text{vec}(Y_t - \widehat{U}_t^j, Y_{t-1}, ..., Y_{t-p}, \widehat{U}_{t-1}^j, ..., \widehat{U}_{t-q}^j)$, $X_{\infty,t}^j = \text{vec}(Y_t - U_t^j, Y_{t-1}, ..., Y_{t-p}, U_{t-1}^j, ..., U_{t-q}^j)$; $\bar{q} = \max\{p, q\}$; and the $(n \times 1)$ vectors $\widehat{\beta}^j$ and β^j stack the n free parameters in the VARMA(p,q) model obtained from the sample and population mappings, respectively. Note that $\widehat{N}_T\left(\widehat{\beta}^j\right)$ and its population counterpart map from $\mathbb{R}^n$ to $\mathbb{R}^n$. The sample and population mappings differ in two important ways. First, the population mapping is a deterministic function, whereas its sample counterpart is stochastic. Second, $\widehat{U}_t^j$ is obtained recursively by

$$\widehat{U}_t^j = Y_t - \left[\widehat{A}_0^j\right]^{-1}\left[\widehat{A}_1^j Y_{t-1} - ... - \widehat{A}_p^j Y_{t-p} - \widehat{M}_1^j \widehat{U}_{t-1}^j - ... - \widehat{M}_q^j \widehat{U}_{t-q}^j\right], \text{ for } t = 1, ..., T, \quad (15)$$

where $Y_{t-\ell} = \widehat{U}_{t-\ell}^j = 0$ for all $\ell \geq t$, while U_t^j is also computed recursively in the same fashion as (15), but assumes that all the pre-sample values are known, i.e., $Y_{t-\ell}$ and $U_{t-\ell}^j$ are known for all $\ell \geq t$. Note that $N(\beta) = \beta$, which implies that when evaluated on the true vector of parameters, the population mapping maps the vector β to itself. This implies that if the population mapping is an ACM then β is a unique fixed point of $N(\beta)$ (see Dominitz and Sherman, 2005, pg. 841).

For theoretical reasons, such that we can formally handle the effect of initial values when computing the residuals recursively, define the infeasible sample mapping as

$$\breve{\beta}^{j+1} = \breve{N}_T\left(\breve{\beta}^j\right) = \left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}\widetilde{X}_{\infty,t}^j\right]^{-1}\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}Y_t\right]. \quad (16)$$

The infeasible sample mapping in (16) differs from the sample mapping because it is function of U_t^j rather than $\widehat{U}_t^j$. In fact, it is a stochastic version of the population mapping and it will be extensively used when deriving the consistency and the asymptotic normality of the IOLS estimator. Lemma 4 shows that, when evaluated at the same vector of estimates, $\widehat{N}_T\left(\widehat{\beta}^j\right)$ converges uniformly to its infeasible counterpart as $T \longrightarrow \infty$, which implies that setting the starting values to zero in (15) do not matter asymptotically.

We start by deriving the consistency and asymptotic distribution of the IOLS estimator for the univariate ARMA(1,1) model,

$$y_t = \beta_1 y_{t-1} + u_t + \beta_2 u_{t-1}. \quad (17)$$

The population, sample and infeasible mappings nest the univariate ARMA(1,1) in (17) with

$\beta = (\beta_1, \beta_2)'$, $K = 1$, $p = q = 1$ and $R = (0, I_2)'$. We impose the following assumptions:

A.1 (Stability, Invertibility) The model in (17) is stable, invertible, and contains no common factors, i.e., $|\beta_1| < 1$, $|\beta_2| < 1$ and $\beta_1 \neq -\beta_2$.

A.2 (Disturbances) The disturbance u_t in (17) is independent and identically distributed (iid) process with $\mathbb{E}(u_t) = 0$, $Var(u_t) = \sigma_u^2$ and finite fourth moment.

We shall prove the consistency of the IOLS estimator using the framework developed in Dominitz and Sherman (2005). We first show that the population mapping is an ACM and thus has a fixed point (see Lemma 1 in the Appendix). To be more precise, Lemma 1 guarantees that the population mapping is an ACM if $\left|\frac{\beta_1\beta_2}{1+\beta_1\beta_2}\right| < 1$.[8] As in Hannan and McDougall (1988) a sufficient rule for Lemma 1 to hold is $\beta_1\beta_2 > -1/2$. The validity of Lemma 1 is crucial to prove the consistency of the IOLS estimator. We denote ϕ and γ as any vectors of estimates of β that satisfy both Assumption A.1 and the contraction condition in Lemma 1. If $N(\phi)$ is an ACM on $(\mathbb{B}, E_2)$, where E_2 is the Euclidean metric on $\mathbb{R}^2$ and $\mathbb{B}$ is a closed ball centered at β, then $|N(\phi) - N(\gamma)| \leq \kappa |\phi - \gamma|$ holds, with $\gamma, \phi \in \mathbb{B}$ and $\kappa \in [0, 1)$. Moreover, $N(\phi)$ will have a unique fixed point on $(\mathbb{B}, E_2)$, as discussed in Dominitz and Sherman (2005).

To establish the asymptotic distribution of the IOLS estimator, define $\widehat{\beta}$ as the fixed point of sample mapping, such that $\widehat{N}_T\left(\widehat{\beta}\right) = \widehat{\beta}$; $J = [I_2 - V(\beta)]^{-1}$, where $V(\beta) = \left.\frac{\partial N(\beta^j)}{\partial \beta^{j\prime}}\right|_{\beta}$ is the gradient of the population mapping evaluated at the true vector of parameters; and $H = \text{plim}\left[\frac{1}{T}\sum_{t=2}^{T} x_t x_t'\right]$, where $x_t = (y_{t-1}, u_{t-1})'$. Theorem 1 delivers the consistency and asymptotic distribution of the IOLS estimator.

Theorem 1 *Suppose Assumptions A.1 and A.2 hold and* $\left|\frac{\beta_1\beta_2}{1+\beta_1\beta_2}\right| < 1$. *Then,*

i. $\left|\widehat{\beta} - \beta\right| = o_p(1)$ *as* $j,\ T \longrightarrow \infty$;

ii. $\sqrt{T}\left[\widehat{\beta} - \beta\right] \xrightarrow{d} \mathcal{N}(0, \Sigma_\beta)$ *as* $j,\ T \longrightarrow \infty$ *and* $\frac{\ln(T)}{j} = o(1)$.

where $\Sigma_\beta := \sigma_u^2 J H^{-1} J'$.

The proof of Theorem 1 and a closed-form expression for Σ_β are given in the Appendix. Despite $\widehat{\beta}$ in Theorem 1 being the fixed point, we make explicit the requirements that

[8] Figure S.2 in the online Supplement displays the maximum eigenvalue associated with the theoretical gradient for different β's satisfying Assumption A.1. This is a theoretical result computed using (S.23), and it is a necessary but not a sufficient condition for the validity of Theorem 1.

$j \longrightarrow \infty$ and $\frac{\ln(T)}{j} = o(1)$ in the statement of Theorem 1, as these are necessary for both the consistency and asymptotic normality of $\widehat{\beta}$.

The validity of Theorem 1 follows from Lemmas 1, 3, 4, 5, 6, 7 and 8. Lemmas 1 and 3 are specific to the ARMA(1,1) model, whereas the remaining Lemmas nest both the ARMA(1,1) and the general VARMA(p,q) models. The proofs for the Lemmas are relegated to Section S. 2 in the online Supplement. To show the consistency and asymptotic distribution of the IOLS estimator, we require four further conditions additional to Lemma 1: sample and infeasible sample mappings converge uniformly in probability as $T \longrightarrow \infty$ (Lemma 4); the population and sample mappings converge uniformly in probability (Lemma 5); uniform convergence on the gradients of the mappings (Lemma 6); the sample mapping is also an ACM (Lemma 7); and $\sqrt{T}$ convergence of $\widehat{\beta}^j$ to the fixed point of the sample mapping as $j, T \longrightarrow \infty$ and $\frac{\ln(T)}{j} = o(1)$ (Lemma 8).

As discussed in Dominitz and Sherman (2005), the results in Theorem 1 hold for any starting value, provided that $\widehat{\beta}^0 \in \mathbb{B}$: Note also that $\widehat{\beta}^j$ converges uniformly in probability to $\widehat{\beta}$. This result is formalized in Lemma 7. In the particular case of the ARMA(1,1) model, it is enough to choose initial estimates that fulfill Assumption A.1 and the contraction condition in Lemma 1, which turns out to be very simple. It is also relevant to note that item (i.) in Theorem 1 holds for any rate of $j \longrightarrow \infty$. Section S. 4 in the online Supplement discusses the asymptotic efficiency loss attached to the IOLS estimator compared to the efficient MLE estimator.

We now turn our attention to extend the results of Theorem 1 to the much more complex case of strong VARMA(p,q) models. We derive our results in a generic way such that the univariate specification is also encompassed. We start by generalizing the set of Assumptions in the univariate ARMA(1,1) model and summarize them in the set of Assumptions B below.

B.1 (Stability, Invertibility and Uniqueness) Let Y_t be a stable and invertible K-dimensional VARMA(p,q) process. Moreover, assume Y_t is uniquely identified and expressed in Echelon form as in (13) with known Kronecker indices.

B.2 (Disturbances) The disturbance U_t in (13) is independent and identically distributed (iid) process with $\mathbb{E}(U_t) = 0$, $Var(U_t) = \Sigma_u$ and finite fourth moment.

B.3 (Contraction and Stochastic Equicontinuity) Define the $(n \times n)$ infeasible sample gradient as $\breve{V}_T\left(\breve{\beta}^j\right) = \frac{\partial \breve{N}_T(\breve{\beta}^j)}{\partial \breve{\beta}^{j\prime}}$ and its population counterpart as $V\left(\beta^j\right) = \frac{\partial N(\beta^j)}{\partial \beta^{j\prime}}$;

and $\mathbb{B}$ as the closed ball centered at β satisfying invertibility and stability conditions in Assumption B.1. Assume that the following hold:

i. The maximum eigenvalue associated with $V(\beta) = \left.\frac{\partial N(\beta^j)}{\partial \beta^{j\prime}}\right|_{\beta}$ is smaller than one in absolute value.

ii. $\sup_{\phi\in\mathbb{B}} \|\breve{V}_T(\phi)\| = O_p(1)$, with $\phi \in \mathbb{B}$.

Assumption B.1 provides the general regularity conditions governing the VARMA(p,q) model. Item i. in Assumption B.3 is enough to guarantee that the IOLS mapping is an ACM on $(\mathbb{B}, E_n)$ and therefore determines the existence of the fixed point and ultimately the asymptotic results governing the IOLS estimator. Lemma 2 provides the sample counterpart of this result, making therefore possible to verify whether the sample mapping is an ACM. Albeit the result in Lemma 2 is computationally easy to obtain, we could not pin down the eigenvalues of the population counterpart of Lemma 2 solely as function of the parameters matrices eigenvalues. In fact, numerical simulation indicates that the maximum eigenvalue of $V(\beta)$ depends on the elements of the parameters matrices in (13) rather than on their eigenvalues. Similarly to the univariate case, we note that the population mapping is not an ACM if some the eigenvalues from the AR and MA components have opposite signs and are close to one in absolute value.

Item ii. in Assumption B.3 generalizes the result in Lemma 3 for the multivariate specification. Assumptions B.1, B.2 and B.3 feed directly into Lemmas 4, 5, 6, 7 and 8. To establish the asymptotic distribution of the IOLS estimator for the general strong VARMA(p,q) model, we adopt similar steps as in Theorem 1. Define $Z = [R'(H \otimes I_K) R]^{-1}$, $H = \text{plim}\left[\frac{1}{T}\sum_{t=\bar{q}+1}^{T} X_t X_t'\right]$ and $J = [I_n - V(\beta)]^{-1}$. Theorem 2 delivers the consistency and asymptotic distribution of the IOLS estimator for the general VARMA(p,q) model.

Theorem 2 *Suppose Assumptions B.1, B.2 and B.3 hold. Then,*

i. $\left|\widehat{\beta} - \beta\right| = o_p(1)$ *as* $j,\ T \longrightarrow \infty$;

ii. $\sqrt{T}\left[\widehat{\beta} - \beta\right] \xrightarrow{d} \mathcal{N}(0, JZR'(H \otimes \Sigma_u) RZ'J')$ *as* $j,\ T \longrightarrow \infty$ *and* $\frac{\ln(T)}{j} = o(1)$.

Lemma 2 and estimates of β can be used to compute the empirical counterparts of $V(\beta)$, H, Z and Σ_u, yielding a feasible estimate of the asymptotic variance.

The practical implication of the violation of item i. in Assumption B.3 is that the IOLS algorithm does not converge even for large T.[9] If this is the case, the asymptotic results derived in this section cannot be implemented. Moreover, we note from our simulations that if the IOLS algorithm converges, the contraction property assumption is satisfied. The opposite, however, is not true, because it is possible to have a DGP satisfying item i. in Assumption B.3 and the IOLS algorithm still fails to converge. This follows because Lemma 7 holds only asymptotically, and convergence in finite sample requires that both the population and sample mappings are ACMs, with the latter holding on every iteration. The Monte Carlo study shows that when the population mapping is an ACM (item i. in Assumption B.3 holds), convergence rates increase monotonically with T.[10]

The asymptotic theory we adopt to derive the consistency and asymptotic normality of the IOLS estimator is flexible enough to allow an extension to the case of weak ARMA(1,1) and weak VARMA(p,q) models. To obtain these results, it is necessary to replace B.2 (A.2 in the univariate case) in such a way we move away from a strong VARMA(p,q) and ARMA(1,1) models to their weak counterpart. Weak VARMA and ARMA processes are characterized by having innovations that are linear projections, being therefore only uncorrelated and potentially non-independent processes. To address this issue, we follow Francq, Roy, and Zakoian (2005), Francq and Zakoian (1998) and Dufour and Pelletier (2014) and replace the *iid* innovations with mixing conditions satisfying the weak VARMA and ARMA definitions (Assumption B.2a). It is important to highlight that these mixing conditions are valid for a wide range of nonlinear models that allow weak ARMA representations (see Francq, Roy, and Zakoian (2005), Francq and Zakoian (1998) and Francq and Zakoian (2005)), and imply a finite fourth moment as required by our IOLS framework.

B.2a Disturbances - strong mixing Let U_t be a $(K \times 1)$ vector of innovations with $K \geq 1$. The disturbances U_t are strictly stationary with $\mathbb{E}(U_t) = 0$, $Var(U_t) = \Sigma_u$, $Cov(U_{t-i}, U_{t-j}) = 0$ for all $i \neq j$ and satisfy the following two conditions:

i. $\mathbb{E}|U_t|^{4+2\nu} < \infty$,

ii. $\sum_{\kappa=0}^{\infty} \{\alpha_u(\kappa)\}^{\nu/(2+\nu)} < \infty$, for some $\nu > 0$,

[9] Using the ARMA(1,1) specification as an example, violation of Assumption B.3 item i. is equivalent to have β_1 and β_2 so that $\beta_1\beta_2 \leq -1/2$. If this is the case, the IOLS does not convergence even for large T, as neither the population nor the sample mappings are ACMs.

[10] Figure S.3 in the online Supplement displays finite sample convergence rates (heat-map) of the IOLS estimator for ARMA(1,1) models. We show that convergence rates improve dramatically when sample size increases from $T = 100$ to $T = 10,000$ in the entire set of parameters satisfying Lemma 1.

where $\alpha_u(l) = \sup_{\substack{\mathcal{D}\in\sigma(U_i, i\geq t+l)\\ \mathcal{C}\in\sigma(U_i, i\leq t),}} |Pr(\mathcal{C}\bigcap\mathcal{D}) - Pr(\mathcal{C})\,Pr(\mathcal{D})|$ are strong mixing coefficients of order $l \geq 1$, with $\sigma(U_i, i \leq t)$ and $\sigma(U_i, i \geq t+l)$ being the σ-fields generated by $\{U_i : i \leq t\}$ and $\{U_i : i \geq t+l\}$, respectively.

Note that Assumption B.2a nests the weak ARMA(1,1) and the general weak VARMA(p,q) models, $K = 1$ and $K > 1$, respectively. Corollary 1 summarizes the theoretical properties of the IOLS estimator for weak VARMA(p,q) models.

Corollary 1 *Suppose Assumptions B.1, B.2a and B.3 hold. Then, item i. in Theorem 2 holds and*

i. $\sqrt{T}\left[\widehat{\beta} - \beta\right] \xrightarrow{d} \mathcal{N}\left(0, JZ\mathcal{I}Z'J'\right)$ *as* j, $T \longrightarrow \infty$ *and* $\frac{\ln(T)}{j} = o(1)$,
where $\mathcal{I} = \sum_{\ell=-\infty}^{\infty}\mathbb{E}\left\{\left[R'\left(X_t \otimes I_K\right)U_t\right]\left[R'\left(X_{t-\ell} \otimes I_K\right)U_{t-\ell}\right]'\right\}$.

Following the work of Francq and Zakoian (1998) and Dufour and Pelletier (2014), Corollary 1 makes use of the central limit theorem of Ibragimov (1962) that encompasses strong mixing processes such as the one in Assumption B.2a. This yields an asymptotic variance that is a function of $\mathcal{I}$ rather than the usual $R'\left(H \otimes \Sigma_u\right)R$ term. As in Theorem 2, Lemma 2 and $\widehat{\beta}$ can be used to compute the finite sample counterparts of the J, Z and $\mathcal{I}$. Specifically, $\mathcal{I}$ can be consistently estimated by the Newey-West covariance estimator,

$$\widehat{\mathcal{I}} = \frac{1}{T-\bar{q}}\sum_{\ell=-m_T}^{m_T}\left[1 - \frac{|\ell|}{m_T+1}\right]\sum_{t=\bar{q}+1+|\ell|}^{T}\left\{\left[R'\left(\widehat{X}_t \otimes I_K\right)\widehat{U}_t\right] \times \left[R'\left(\widehat{X}_{t-\ell} \otimes I_K\right)\widehat{U}_{t-\ell}\right]'\right\}, \tag{18}$$

where $m_T^4/T \to 0$ with $T, m_T \to \infty$. Therefore, from a practitioner's point of view, it suffices to adopt the well known Newey-West estimator as in (18), when constructing the asymptotic variance of the IOLS estimator, to handle weak VARMA(p,q) models.

Remark 1: Corollary 1 nests the ARMA(1,1) model with $\beta = (\beta_1, \beta_2)'$, $K = p = q = 1$ and $R = (0, I_2)'$. It also holds under weaker assumptions for the ARMA(1,1) specification. Specifically, item i. in Assumption B.3 can be replaced by $\left|\frac{\beta_1\beta_2}{1+\beta_1\beta_2}\right| < 1$ as it suffices to guarantee that the population mapping is an ACM; and item ii. in Assumption B.3 is no longer necessary following that Lemma 3 holds in the entire set of parameters satisfying Assumption A.1.

4 Monte Carlo Study

This section provides results that shed light on the finite sample performance of VARMA models estimated using the IOLS methodology. We compare the IOLS estimator with estimators possessing very different asymptotic and computational characteristics. We report results considering five different methods: the MLE, the two-stage (HR) method of Hannan and Rissanen (1982), the three-step (HK) procedure of Hannan and Kavalieris (1984), the two-stage (DJ2) method of Dufour and Jouini (2014), and the multivariate version of the three-step (KP) procedure of Koreisha and Pukkila (1990) as discussed in Koreisha and Pukkila (2004) and Kascha (2012). To broadly analyse and assess the performance of the IOLS estimator, we design simulations covering different sample sizes (ranging from 50 to 1000 observations), system sizes ($K = 3$, $K = 10$, $K = 20$, $K = 40$ and $K = 52$), Kronecker indices, dependencies among the variables, and allow for both strong and weak processes.

We simulate stable, invertible and unique VARMA(1,1) models,

$$A_0 Y_t = A_1 Y_{t-1} + A_0 U_t + M_1 U_{t-1}. \tag{19}$$

Uniqueness is imposed through the Echelon form transformation, which implies $A_0 = M_0$ in (19) and requires a choice of Kronecker indices. We discuss results considering five different DGPs. DGPs I and II set all Kronecker indices to one, which implies that $A_0 = I_K$ and A_1 and M_1 are full matrices. The two DGPs differ with respect to the eigenvalues assigned to the parameter matrices. The eigenvalues in DGP I are constant and equal to 0.5, whereas the eigenvalues in DGP II take positive, negative and near-to-zero values. DGPs III, IV and V impose a similar structure to the Kronecker indices. The first k Kronecker indices are set to one, while the remaining $K-k$ Kronecker indices are set to zero, so that $\mathbf{p} = (p_1, p_2, ..., p_K)'$ with $p_i = 1$ for $i \leq k$ and $p_i = 0$ for all $i > k$. Specifically, DGP III has $k = 1$, while DGP IV and DGP V have $k = 2$ and $k = 3$, respectively. The free parameters in DGPs III, IV and V are based on real data, so that they are chosen as the estimates obtained by fitting VARMA(1,1) models to the dataset studied in Section 5.[11] DGPs III, IV and V are particularly relevant because they reduce dramatically the number of free parameters in (19), while yielding rich dynamics in the MA component of the standard representation of (19).[12]

[11]See Section S. 5 in the online Supplement for a complete description of the DGPs used in this Section.

[12]Because A_0 is an invertible lower triangular matrix and $A_0 \neq I_K$, multiplying (19) by A_0^{-1} yields the

Strong VARMA(1,1) models are obtained by generating U_t with *iid* Gaussian innovations, while U_t in weak VARMA(1,1) models, are generated as in Romano and Thombs, 1996, pg. 591, with $U_t = \prod_{\ell=0}^{m} \varepsilon_{t-\ell}$, where $m > 1$ and ε_t is a zero mean *iid* process with covariance matrix I_K. This procedure yields uncorrelated innovations satisfying the mixing conditions stated in Assumption B.2a. We summarize results for each specification using two measures: MRRMSE and Share. MRRMSE accounts for the mean of the relative root median squared error (RRMSE) measures of all parameters, where RRMSE is the ratio of the root median squared error (RMSE) obtained from a given estimator over the RMSE of the HR estimator. RRMSE measures lower than one indicates that the HR estimator is outperformed by the alternative estimator. Share is the frequency a given estimator returns the lowest RRMSE over all the free parameters.[13] The MRRMSE and Share measures are only computed using the replications that achieved convergence and satisfy Assumption B.1[14]. We discard the initial 500 observations and fix the number of replications to 5,000, unless otherwise stated. This paper reports only a fraction of the entire set the Monte Carlo results. The complete set of tables can be found on Section S. 5 in the online Supplement.

The first set of Monte Carlo simulations addresses the finite sample performance of the IOLS estimator in small sized ($K = 3$) VARMA(1,1) models. We simulate strong and weak VARMA(1,1) models considering three alternative DGPs. DGPs I and II yield 18 free parameters, while DGP III has 6 free parameters. We consider samples of 50, 100, 150, 200 and 400 observations. Table 1 brings the results for the weak VARMA(1,1) models. First, we find that the MLE estimator is dominant in DGP I. This is not a surprising finding, because MLE is known to perform well on specifications where the absolute eigenvalues are bounded away from zero and one (see extensive study in Kascha (2012)). Secondly, a different picture arises when analysing the results from DGP II. In one hand, we find that this specification is numerically more difficult to handle, yielding lower rates of convergence for both the IOLS and MLE estimators. For the particular case of the IOLS, we note that IOLS fails because the invertibility condition is often violated at some iteration. As

standard representation of (19), $Y_t = A_0^{-1}A_1Y_{t-1} + U_t + A_0^{-1}M_1U_{t-1}$, where $A_0^{-1}M_1$ is a full matrix. See discussion in Section S. 3 in the online Supplement

[13] As an example, when $K = 3$ and $\mathbf{p} = (1, 0, 0)'$, there are six free parameters to be estimated. If the IOLS estimator has a Share of 67%, it implies that the IOLS estimator delivered the lowest RRMSE in four out of those six free parameters. This measure is particularly informative when dealing with systems with large number of free parameters.

[14] Throughout this section, we assume an estimator converges if its final estimates satisfy Assumption B.1. Additionally for the IOLS and MLE estimators, convergence also implies numerical convergence of their respective algorithms.

T increases, convergence rates for the IOLS estimator increases. This reflects the fact that Lemma 2 holds when evaluated at the true vector of parameters and Lemma 7 holds only asymptotically. On the other hand, we find that the IOLS estimator is the one which delivers the best results considering both MRRMSE and Share measures. This indicates that if convergence is achieved, the IOLS estimator is able to handle systems with near-to-zero eigenvalues more efficiently than the benchmark MLE estimator. Third, we find that IOLS is very competitive in DGP III. In particular, we show that the IOLS presents the highest Share measure in all sample sizes. More specifically, we find that the IOLS estimator performs particularly well on estimating the two free parameters in A_0.

The last set of simulations investigates the finite sample performance of the IOLS estimator in medium and large sized systems. To this purpose, we simulate strong and weak versions of DGPs III, IV and V with $K = 10$, $K = 20$, $K = 40$ and $K = 52$. These are the system dimensions and the Kronecker indices we further consider in our empirical study. As far our knowledge goes, this is the first study to consider such high dimensional VARMA models in a Monte Carlo study. The sample size and number of replications are set to $T = 400$ and 1000, respectively. These are high dimensional models with free parameters varying from 20 to 312. Table 2 displays the results. We now find that the IOLS estimator presents the best relative performance (in terms of both the MRRMSE and Share measures) for large systems ($K = 40$ and $K = 52$). These findings hold for all the three DGPs and both strong and weak processes. The relative performance of the HK and IOLS estimators are outstanding, with an average improvement with respect to the HR estimator of up to 65%. On average, the IOLS outperforms the HK estimator in 11% (in terms of the MRRMSE measure). Considering systems with $K \leq 20$, we find that the HK estimator is the one that delivers the best performance, while the IOLS estimator is constantly ranked as the second best estimator.[15] Finally, we conclude that the IOLS estimator considerably improves its performance when estimating high dimensional restricted models with $A_0 \neq I_K$, while remaining a feasible alternative (average convergence rate of 97%). These results motivate us to implement the Echelon form transformation in the fashion of DGPs III, IV and V and the IOLS estimators as feasible alternatives to estimate high dimensional VARMA(1,1) models.

Overall, we conclude that the IOLS estimator is a competitive alternative and compares

[15] Table S.8 in the online Supplement shows that the IOLS outperforms the HK estimator for both $K = 10$ and $K = 20$, when estimating weak processes with $T = 200$.

favourable with its competitors in a variety of cases: small sample sizes; small sized systems with near-to-zero eigenvalues; large sized systems with many Kronecker indices set to zero. The MLE and HK estimators also present remarkable performances in terms of RMSE, which is in line with previous studies (see Kascha (2012)).

5 Empirical Application

In this section, we analyze the competitiveness of VARMA models estimated with the IOLS procedure to forecast macroeconomic variables. We forecast three key macroeconomic variables: industrial production (IPS10), interest rate (FYFF), and CPI inflation (PUNEW). We assess VARMA forecast performance under different system dimensions and forecast horizons.

5.1 Data and Setup

We use U.S. monthly data from the Stock and Watson (2006) dataset, which runs from 1959:1 through 2003:12. We do not use all the available series from this dataset; as in Carriero, Kapetanios, and Marcellino (2011), we use 52 macroeconomic variables that represent the main categories of economic indicators. From the 52 selected variables, we work with five system dimensions: $K = 3$, $K = 10$, $K = 20$, $K = 40$, and $K = 52$. We construct four different datasets (one to four) for each system size where $10 \leq K \leq 40$, as a way to assess robustness of the VARMA framework when dealing with different explanatory variables. These results are also very useful to understand the forecast performance of VARMA models using medium sized datasets, as well as to understand the trade off between forecast performance and the estimation cost associated with larger systems. There is no particular rule to select the variables within the entire group of 52; however, we try to keep a balance among the three main categories of data: real economy, money and prices, and financial market.[16] The series are transformed, as in Carriero, Kapetanios, and Marcellino (2011), in such a way that they are approximately stationary. The forecasting exercise is performed in pseudo real time, with a fixed rolling window of 400 observations. All models considered in the exercise are estimated on every window. We perform 115 out-of-sample forecasts considering six different horizons: one- (Hor:1), two- (Hor:2), three- (Hor:3), six- (Hor:6), nine- (Hor:9) and twelve- (Hor:12) steps-ahead.

[16] Table S.9 in the online Supplement reports the details for all datasets.

We compare the different VARMA specifications with four alternative methods: AR(1), VAR(p^*), BVAR and factor models.[17] Factor models summarize a large number of predictors in only a few number of factors. We forecast the three key macroeconomic variables using the two-step procedure discussed in Stock and Watson (2002a) and Stock and Watson (2002b). In the first step, factors ($\{F_t\}_{t=1}^T$) are extracted via principal components, whereas the second step consists of projecting $y_{i,t+l}$ onto $\left(\widehat{F}_t', ..., \widehat{F}_{t-l}', y_{i,t-1}, ..., y_{i,t-r}\right)$, with $l \geq 0$ and $r > 0$. We follow Stock and Watson (2002a) and determine the lag orders by minimizing the Schwarz criterion (SC) criterion. The number of factors is chosen according to the SC and IC_{p3} criteria as in Stock and Watson (2002a), denoted as FM_{SC} and FM_{IC3}, respectively. Factors are computed using only the variables available on the respective dataset. The BVAR framework builds on the idea of applying Bayesian shrinkage via the imposition of prior beliefs on the parameters of a K dimensional stable VAR(p) model. We estimate the BVAR model with the normal-inverted Wishart prior as in Bańbura, Giannone, and Reichlin (2010).[18] It is important to note that we follow Bańbura, Giannone, and Reichlin (2010) and adjust the prior to accommodate the fact that our variables are approximately stationary. The hyperparameter φ (tightness parameter) plays a crucial role on the amount of shrinkage we impose on the parameter estimates and hence on the forecast performance of the BVAR models. When $\varphi = 0$, the prior is imposed exactly, while $\varphi = \infty$ yields the standard OLS estimates. We report results considering three BVAR specifications. The first set of results, denoted as BVAR_{SC}, is obtained by setting φ to the value which minimizes the SC criterion over a grid of $\varphi \in (2.0e-5, 0.0005, 0.002, 0.008, 0.018, 0.072, 0.2, 1, 500)$.[19] The second set of results, denoted as $\text{BVAR}_{0.2}$, sets $\varphi = 0.2$. This is the default choice of hyperparameter in the package Regression Analysis of Time Series (RATS) and it has been used as a benchmark model in Carriero, Kapetanios, and Marcellino (2011). The third specification, BVAR_{opt}, follows from Bańbura, Giannone, and Reichlin (2010) and chooses $\varphi \in (2.0e-5, 0.0005, 0.002, 0.008, 0.018, 0.072, 0.2, 1, 500)$ as the hyperparameter which minimizes the in-sample one-step-ahead root mean squared forecast error in the last 24 months of the sample. For all the different specifications, we choose the lag length that

[17] The lag length p^* in the VAR(p^*) specification is obtained by minimizing the AIC criterion.

[18] See the online Supplement (Section S. 6) for an extended discussion on the BVAR framework implemented in this section.

[19] $\varphi \in (2.0e-5, 0.0005, 0.002, 0.008, 0.018, 0.072, 0.2, 1, 500)$ follows from the grid search in Carriero, Kapetanios, and Marcellino (2011) and it is broad enough to include both very tight ($\varphi = 2.0e-5$) and loose ($\varphi = 500$) hyperparameters.

minimizes the SC criterion.[20] We grid search over φ and the optimal lag length on every rolling window.

Up to this moment, we have assumed that the Kronecker indices are all known, which implies that any general VARMA model can be written in Echelon form by applying the procedure described by equations (4), (5), (6) and (7). When one is dealing with empirical data, however, the true DGP is unknown as, consequently, are the Kronecker indices. We determine the Kronecker indices for the VARMA specifications using two strategies. The first one uses the Kronecker indices to impose rank reduction on the parameter matrices of the VARMA(1,1) model as in the DGPs III, IV and V discussed in the Monte Carlo section. This follows the work of Carriero, Kapetanios, and Marcellino (2011), who show that reduced rank VAR (RRVAR) models perform well when forecasting macroeconomic variables using large datasets. By specifying the Kronecker indices as $\mathbf{p} = (1, 0, ..., 0)'$, $\mathbf{p} = (1, 1, 0, ..., 0)'$, $\mathbf{p} = (1, 1, 1, 0, ..., 0)'$, the rank of both $A_0^{-1}A_1$ and $A_0^{-1}M_1$ reduces to one, two and three, respectively. We denote these specifications as $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$, respectively. Note that the restrictions and number of free parameters in $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ are analogous to DGPs III, IV and V, respectively. To assess how these three different specifications fit the data, we report two different SC criteria. The first one, denoted as SC_K, is the standard SC criterion that is computed with the entire $(K \times K)$ covariance matrix of the residuals. The second criterion, denoted as SC_3, computes the SC criterion using only the (3×3) upper block of the residuals covariance matrix, as this upper block contains the covariance matrix associated with the three key macroeconomic variables. The SC_3 criterion, therefore, is a measure of fit that is solely related to the variables that we are ultimately interested in. The second strategy adopts the Hannan-Kavalieris algorithm (denoted as $\mathbf{p}_{HK}$), that consists of choosing Kronecker indices that minimize the SC_K criterion, as this information criterion delivers consistent estimates of the Kronecker indices when Assumption A.1 and A.2 are satisfied.[21] The Hannan-Kavalieris procedure is implemented on every rolling window for $K = 3$ and $K = 10$. For medium and large sized datasets ($K = 20$, $K = 40$ and $K = 52$), we apply the Hannan-Kavalieris algorithm on the first rolling window and carry the estimated Kronecker indices to the subsequent ones. VARMA models are estimated with the IOLS estimator discussed in Section 2.[22]

[20] The maximum lag length is set to be 15, 8 and 6 for $K \leq 10$, $20 \leq K \leq 40$ and $K = 52$, respectively.

[21] See Lütkepohl, 2007, pg. 503 and Section S. 6.2 in the online Supplement for a complete description of the Hannan-Kavalieris procedure.

[22] As discussed in Section 2, if the IOLS estimator does not converge, we adopt the consistent initial

We compare different models using the out-of-sample relative mean squared forecast error (RelMSFE) computed using the AR(1) as a benchmark. We choose to have the AR(1) as a benchmark, because this makes easy the comparison across the different datasets. We assess the predictive accuracy among the different models using the Diebold and Mariano (1995) test. The use of this test is justified given our focus on forecasts obtained through rolling windows.

5.2 Results

We organize the results as follows. Table 3 reports a summary of the forecast results computed across the three key macroeconomic variables and datasets. Specifically, Table 3 presents, for each system size, five panels. The first panel reports the frequency (in percentage points) for which at least one of the VARMA specifications ($\mathbf{p}_{100}$, $\mathbf{p}_{110}$, $\mathbf{p}_{111}$ and $\mathbf{p}_{HK}$) outperforms (delivers the lowest RelMSFE measures) the assigned group of competitors in a given forecast horizon.[23] Similarly, the second, third, fourth and fifth panels display the frequencies for which the $\mathbf{p}_{100}$, $\mathbf{p}_{110}$, $\mathbf{p}_{111}$ and $\mathbf{p}_{HK}$ specifications, respectively, outperform the assigned group of competitors. We consider four groups of competitors: AR, VAR, FM and BVAR. The AR group collects the AR(1) specification; VAR contains the VAR(p^*) model; FM gathers the different factor model specifications, namely the FM_{SC} and FM_{IC3}; and the BVAR collects the three BVAR specifications: BVAR_{SC}, $\text{BVAR}_{0.2}$ and BVAR_{opt}. The online Supplement (Tables S.10, S.11a, S.11b, S.12a, S.12b, S.13a and S.13b) has the detailed results, so that it is possible to assess the forecast performance of VARMA models up to the level of the key macroeconomic variables. Finally, Table 4 compares the performance of the IOLS estimator with the DJ2, HK and KP estimators. We report the frequency (in percentage points) for which the IOLS estimator outperforms these alternative VARMA estimators in a given forecast horizon.

Starting with $K = 3$, we find that VARMA models largely outperform the AR, VAR, FM and the BVAR groups up to the sixth-step-ahead forecast, (see Table 3). Specifically, we find that the $\mathbf{p}_{100}$ and $\mathbf{p}_{110}$ specifications are the ones which deliver the best results. In fact, these are the specifications for which the IOLS estimator converges in 100% of the rolling windows (see Table S.10). Taking into account all forecast horizons, VARMA models

estimates: the two-stage HR estimator.

[23]Percentages are computed across the three key macroeconomic variables and the four different datasets discussed in page 22 and specified in details in Table S.9.

deliver the best forecast in 72% of the cases.

We now turn our attention to the results obtained using the entire dataset ($K = 52$). When comparing the performance of VARMA models with the BVAR group, we find that the former delivers lower RelMSFE measures in 72% of the cases. This is a strong result in favour of VARMA models, because BVAR specifications are known to be very competitive when forecasting with large datasets. Overall, factor models deliver the best overall performance (see Table 3). Among the VARMA specifications, the $\mathbf{p}_{100}$ and $\mathbf{p}_{110}$ specifications deliver the best results. They do particularly well on forecasting the PUNEW variable (see Table S.10). We find that the IOLS estimator presents convergence rates of 100% for the $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ specifications. This finding reinforces two important aspects: using the Echelon form transformation in the fashion of DGPs III, IV and V is a powerful tool to deal with high dimensional models; and that IOLS estimator is particularly suitable to estimate high-dimensional VARMA models, as our Monte Carlo simulations suggest.

Table S.10 also allows us to compare the performance of the VARMA models estimated with $K = 3$ with factor models and BVAR specifications that use all the 52 variables. We find that VARMA models estimated using small sized datasets produce results as accurate as the factor models and BVAR computed with the entire dataset. Specifically, we find that small sized VARMA models outperform these factor models and BVAR specifications in 67% of the cases (Hor:1 - Hor:6). This empirical finding reinforces the theoretical result that a latent dynamic factor model yields observed variables that follow a VARMA specification (see Dufour and Stevanović (2013)).

Comparing the forecast performance of VARMA models estimated with the IOLS and the alternative VARMA estimators, we find that IOLS largely outperforms (presents values than 50%) all the competitors (see top-left and bottom panels in Table 4). Specifically, the IOLS estimator delivers the best VARMA based forecast in 92% and 90% of the horizons for $K = 3$ and $K = 52$, respectively. We thus conclude from Table 4 that the highly competitive results in Tables 3 and S.10 are due to the VARMA framework combined with the use of the IOLS estimator.

We now summarize the results for the medium and large datasets, ($K = 10$, $K = 20$ and $K = 40$). We start by discussing the results for $K = 10$. For short horizons (Hor:1 - Hor:6), VARMA models deliver the lowest RelMSFE among all groups in 67% of the cases. Compared with the BVAR group, VARMA models remain dominant, delivering more accurate forecasts in 73% of the cases (Hor:1 - Hor:6). Moreover, we find that the $\mathbf{p}_{100}$ specification is the one that usually delivers the lowest RelMSFE measures among the VARMA specifications. This is also the specification with the lowest SC_3 criterion in all datasets, which indicates that choosing the Kronecker indices that minimizes the SC_3 criterion may pay off in terms of forecast accuracy. With regard to the estimation of VARMA models, the IOLS algorithm works well, failing to converge in only 2% of the cases. Its relative performance with respect to the DJ2, HK and KP estimators is also positive. Considering the $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ specifications, the IOLS estimator outperforms its linear competitors in 80% of the horizons (top-right panel in Table 4).

We now discuss the results for $K = 20$. Overall, VARMA models are very competitive in the short horizons (Hor:1-Hor:6), outperforming the AR, VAR, FM and BVAR groups in 88%, 88%, 77% and 71% of the cases, respectively. Results are also stable across the different datasets, showing robustness of the VARMA framework. Moreover, we find that the $\mathbf{p}_{100}$ specification is the one that tends to outperform the other VARMA specifications, indicating that implementing rank reduction on VARMA models improves forecast accuracy, as it does with standard VAR models (Carriero, Kapetanios, and Marcellino (2011)). Additionally, the $\mathbf{p}_{100}$ specification is the one that minimizes the SC_3 criterion in all datasets. Rates of convergence exceed 91% for the IOLS algorithm, suggesting that as long as we restrict the number of free parameters in the fashion of DGPs III, IV and V, estimation of high-dimensional VARMA models is no longer an issue. When comparing to the DJ2, KP and HK estimators, the IOLS delivers most accurate forecasts in 54% of the cases (see Table 4). The strong performance of the HK estimator goes in line with our Monte Carlo results.

Considering the case of large datasets ($K = 40$), we find that VARMA models stay competitive on forecasting shorter horizons. Results across the different datasets remain stable. The performance of the $\mathbf{p}_{HK}$ specification deteriorates compared with the other three less complex specifications, confirming that overparameterization considerably reduces the forecast performance of VARMA models in large systems.[24] By increasing the system size from $K = 20$ to $K = 40$, the RelMSFE measures obtained using VARMA specifications remain stable. This suggests that the potential gains from adding extra information (increasing the number of variables in the system) are offset by potentially less efficient estimates ("curse of dimensionality"). Factor models improve their performance compared to systems with

[24]The Hannan-Kavalieris procedure yields $n = 546$, $n = 235$, $n = 909$ and $n = 235$ free parameters, for datasets 1,2, 3, and 4, respectively.

$K = 20$, confirming that these methods are particularly good on forecasting using many predictors. On the comparison between VARMA and BVAR models, VARMA models remain in a better position, delivering more accurate forecasts in 75% of the cases (Hor:1 - Hor:6). The IOLS estimator remains a robust alternative, achieving convergence in 80% of the rolling windows. The IOLS estimator continues outperforming the DJ2, HK and HP estimators, delivering more accurate forecasts in 93% of the horizons. This goes in line with our Monte Carlo simulations, which shows that the IOLS estimator delivers an outstanding performance in large sized system, ($K = 40$ and $K = 52$).

To sum up the results of this section, VARMA models estimated with the IOLS estimator are generally very competitive and able to beat the three most prominent competitors in this type of study: AR(1), factor models and BVAR models. This finding is especially present for the $\mathbf{p}_{100}$ specification, which usually minimizes the SC_3 criterion. This result supports previous findings in the literature[25], which conclude that imposing restrictions on the parameter matrices contributes to improve forecast accuracy when dealing with medium and large sized datasets. VARMA results are also stable across the different datasets and Kronecker indices specifications, indicating that the framework adopted is fairly robust. Considering all system sizes and datasets, VARMA specifications deliver the lowest RelMSFE measures for the one-, two-, three-, and six-month-ahead forecast in 54% of the cases, indicating that VARMA models are indeed strong candidates to forecast key macroeconomic variables using small, medium and large sized datasets. It is particularly relevant to highlight the performance of VARMA models relative to the BVAR models. We find that VARMA models systematically deliver more accurate forecasts considering small, medium and large datasets. Finally, the IOLS estimator shows to be a valid alternative to deal with large and complex VARMA systems, presenting convergence rates exceeding 88%. Moreover, the IOLS estimator compares favourable with the main linear competitors, delivering more accurate forecasts in 79% of the cases.

6 Conclusion

This paper addresses the issue of modelling and forecasting key macroeconomic variables using rich (small, medium and large sized) datasets. We propose the use of VARMA(1,1)

[25]See Carriero, Kapetanios, and Marcellino (2011), De Mol, Giannone, and Reichlin (2008) and Bańbura, Giannone, and Reichlin (2010).

models as a feasible framework for this task. We overcome the natural difficulties in estimating medium- and high-dimensional VARMA models with the MLE framework by adopting the IOLS estimator.

We establish the consistency and asymptotic distribution for the IOLS estimator by considering the general strong and weak VARMA(p,q) models. It is also important to point out that our theoretical results are obtained under weak assumptions that are compatible with the QMLE estimator. The extensive Monte Carlo study shows that the IOLS estimator is feasible and consistent in high-dimensional systems. Furthermore, we also report results showing that the IOLS estimator outperforms the MLE and other linear estimators, in terms of mean squared error, in a variety of scenarios: when T is small; disturbances are weak; near-to-zero eigenvalues; and high-dimensional models ($K = 40$ and $K = 52$). The empirical results show that VARMA(1,1) models perform better than AR(1), VAR, BVAR and factor models for different system sizes and datasets. We find that VARMA(1,1) models estimated with the IOLS estimator are very competitive at forecasting short horizons (one-, two-, three- and six-month-ahead horizons) in small ($K = 3$), medium ($K = 10$ and $K = 20$) and large ($K = 40$ and $K = 52$) sized datasets. In particular, VARMA(1,1) models with rank of parameter matrices equal to one ($\mathbf{p}_{100}$) is the one that produces the most accurate results among all VARMA specifications. We find that small sized VARMA models ($K = 3$) deliver forecasts that are as accurate as the ones obtained with factor models using the entire dataset ($K = 52$). This last empirical finding reinforces the theoretical justification for the use of VARMA models, as discussed in Dufour and Stevanović (2013).

References

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

Athanasopoulos, G., and F. Vahid (2008): "A complete VARMA modelling methodology based on scalar components," *Journal of Time Series Analysis*, 29, 533–554.

Bańbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92.

Bernanke, B., J. Boivin, and P. S. Eliasz (2005): "Measuring the Effects of Monetary

Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach," *The Quarterly Journal of Economics*, 120(1), 387–422.

Blanchard, O., and D. Quah (1989): "The dynamic effects of aggregate demand and supply disturbances," *American Economic Review*, 79, 655–673.

Carriero, A., G. Kapetanios, and M. Marcellino (2011): "Forecasting large datasets with Bayesian reduced rank multivariate models," *Journal of Applied Econometrics*, 26, 735–761.

Chan, J. C., E. Eisenstat, and G. Koop (2016): "Large Bayesian VARMAs," *Journal of Econometrics*, 192(2), 374 – 390, Innovations in Multiple Time Series Analysis.

De Mol, C., D. Giannone, and L. Reichlin (2008): "Forecasting Using a Large Number of Predictors: Is Bayesian Regression a Valid Alternative to Principal Components?," *Journal of Econometrics*, 146, 318–328.

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

Dominitz, J., and R. Sherman (2005): "Some Convergence Theory for Iterative Estimation Procedures with an Application to Semiparametric Estimation," *Econometric Theory*, 21, 838–863.

Dufour, J.-M., and T. Jouini (2005): "Asymptotic Distribution of a Simple Linear Estimator for Varma Models in Echelon Form," in *Statistical Modeling and Analysis for Complex Data Problems*, ed. by P. Duchesne, and B. RMillard, pp. 209–240. Springer US.

——— (2014): "Asymptotic distributions for quasi-efficient estimators in echelon VARMA models," *Computational Statistics & Data Analysis*, 73, 69–86.

Dufour, J.-M., and D. Pelletier (2014): "Practical methods for modelling weak VARMA processes: identification, estimation and specification with a macroeconomic application," Manuscript.

Dufour, J.-M., and D. Stevanović (2013): "Factor-Augmented VARMA Models With Macroeconomic Applications," *Journal of Business & Economic Statistics*, 4, 491–506.

FORNI, M., M. HALLIN, M. LIPPI, AND L. REICHLIN (2000): "The Generalised Dynamic Factor Model: Identification and Estimation," *Review of Econometrics and Statistics*, 82(4), 540–554.

FRANCQ, C., R. ROY, AND J.-M. ZAKOIAN (2005): "Diagnostic Checking in ARMA Models with Uncorrelated Errors," *Journal of the American Statistical Association*, 100, 532–544.

FRANCQ, C., AND J.-M. ZAKOIAN (1998): "Estimating linear representations of nonlinear processes," *Journal of Statistical Planning and Inference*, 68, 145–165.

——— (2005): "Recent Results for Linear Time Series Models with Non Independent Innovations," in *Statistical Modeling and Analysis for Complex Data Problems*, ed. by P. Duchesne, and B. RMillard, pp. 241–265. Springer US.

GOURIEROUX, C., AND A. MONFORT (2015): "Revisiting Identification and estimation in Structural VARMA Models," Discussion paper, CREST.

HANNAN, E. J., AND L. KAVALIERIS (1984): "Multivariate Linear Time Series Models". Advances in Applied Probability," *Advances in Applied Probability*, 16(3), 492–561.

HANNAN, E. J., AND A. J. MCDOUGALL (1988): "Regression Procedures for ARMA Estimation," *Journal of the American Statistical Association*, 83, 490–498.

HANNAN, E. J., AND J. RISSANEN (1982): "Recursive estimation of mixed autoregressive-moving average order," *Biometrika*, 69, 81–94.

IBRAGIMOV, A. (1962): "Some Limit Theorems for Stationary Processes," *Theory of Probability and its Applications*, 7, 349–382.

KAPETANIOS, G. (2003): "A Note on an Iterative Least Squares Estimation Method for ARMA and VARMA models," *Economics Letters*, 79(3), 305–312.

KASCHA, C. (2012): "A Comparison of Estimation Methods for Vector Autoregressive Moving-Average Models," *Econometric Reviews*, 31, 297–324.

KOREISHA, S., AND T. PUKKILA (1990): "A generalized least squares approach for estimation of autoregressive moving-average models," *Journal of Time Series Analysis*, 11, 139–151.

——— (2004): "The specification of vector autoregressive Monving average models," *Journal of Statistical Computation and Simulation*, 74, 547–565.

Lütkepohl, H. (2007): *New Introduction to Multiple Time Series Analysis.* Springer-Verlag.

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

Ng, S., and P. Perron (1995): "Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag," *Journal of the American Statistical Association*, 90, 268–281.

Romano, J. P., and L. A. Thombs (1996): "Inference For Autocorrelations Under Weak Assumptions," *Journal of the American Statistical Association*, 91, 590–600.

Rubio-Ramrez, J. F., D. F. Waggoner, and T. Zha (2010): "Structural Vector Autoregressions: Theory of Identification and Algorithms for Inference," *The Review of Economic Studies*, 77, 665–696.

Spliid, H. (1983): "A Fast Estimation Method for the Vector Autoregressive Moving Average Model With Exogenous Variables," *Journal of the American Statistical Association*, 78, 843–849.

Stock, J. H., and M. W. Watson (2002a): "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97(460), 1167–1179.

Stock, J. H., and M. W. Watson (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economics Statistics*, 20, 147–162.

Stock, J. H., and M. W. Watson (2006): "Chapter 10 Forecasting with Many Predictors," vol. 1 of *Handbook of Economic Forecasting*, pp. 515 – 554. Elsevier.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), pp. 267–288.

Uhlig, H. (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure," *Journal of Monetary Economics*, 52(2), 381 – 419.

7 Appendix

Lemma 1 *Suppose Assumptions A.1, A.2 hold and* $\left|\frac{\beta_1\beta_2}{1+\beta_1\beta_2}\right| < 1$*. Then, there exists an open ball centered at* β *with closure* $\mathbb{B}$*, such that the mapping* $N(\phi)$ *is an ACM on* $(\mathbb{B}, E_2)$*, with* $\phi \in \mathbb{B}$ *and* E_2 *denoting the Euclidean metric on* $\mathbb{R}^2$*. Furthermore,* $V(\beta) = \left.\frac{\partial N(\beta^j)}{\partial \beta^{j\prime}}\right|_{\beta}$ *is given by:*

$$V(\beta) = \left.\frac{\partial N(\beta^j)}{\partial \beta^{j\prime}}\right|_{\beta} = \begin{pmatrix} \frac{\beta_2}{\beta_1+\beta_2} & \frac{\beta_2(1-\beta_1^2)}{(\beta_1+\beta_2)(1+\beta_1\beta_2)} \\ \frac{-\beta_2}{\beta_1+\beta_2} & \frac{-\beta_2(1-\beta_1^2)}{(\beta_1+\beta_2)(1+\beta_1\beta_2)} \end{pmatrix}.$$

Proof. See online Supplement.

Lemma 2 *Assume Assumptions B.1 and B.2a hold, then* $\widehat{V}_T\left(\widehat{\beta}^j\right) = \frac{\partial \widehat{N}_T(\widehat{\beta}^j)}{\partial \widehat{\beta}^{j\prime}}$ *is given by:*

$$\begin{aligned}\widehat{V}_T\left(\widehat{\beta}^j\right) = \Bigg\{ & \left[I_1 \otimes Z^{j^{-1}}\right] \frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \Bigg\{ \left(Y_t' \otimes I_n\right)\left(I_K \otimes R'\right) \times \\ & \left[\left(I_1 \otimes \mathbb{K}_{K,f} \otimes I_K\right)\left(I_f \otimes vec\left(I_K\right)\right)\right] \frac{\partial vec\left(\widehat{X}_t^j\right)}{\partial \widehat{\beta}^{j\prime}} \Bigg\} + \\ & \Bigg\{ \left[\left(\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \tilde{X}_t^{j\prime} Y_t\right)' \otimes I_n\right] \left[-\left(Z^j\right)^{-1} \otimes \left(Z^j\right)^{-1}\right] \times \\ & \left[\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \Bigg\{ \left[\left(I_{n^2} + \mathbb{K}_{n,n}\right)\left(I_n \otimes \tilde{X}_t^{j\prime}\right)\left(R' \otimes I_K\right) \times \right.\right. \\ & \left.\left. \left[\left(I_f \otimes \mathbb{K}_{K,1} \otimes I_K\right)\left(I_f \otimes vec\left(I_K\right)\right)\right] \frac{\partial vec\left(\widehat{X}_t^{j\prime}\right)}{\partial \widehat{\beta}^{j\prime}} \Bigg\}\right]\right]\Bigg\}\end{aligned} \tag{20}$$

where $\frac{\partial vec(\widehat{X}_t^j)}{\partial \widehat{\beta}^{j\prime}} = \frac{\partial vec(\widehat{X}_t^{j\prime})}{\partial \widehat{\beta}^{j\prime}}$ *are given in (S.43) and (S.44),* $\mathbb{K}$ *is the commutation matrix,* $f = K(p+q+1)$*, and* $\bar{q} = \max\{p, q\}$*.*

Proof. See online Supplement.

Proof of Theorem 1: Denote $\phi = (\phi_1, \phi_2)'$, with $\phi \in \mathbb{B}$ as a vector collecting the parameter estimates of the ARMA(1,1) model, such that Assumption A.1 and $\left|\frac{\phi_1\phi_2}{1+\phi_1\phi_2}\right| < 1$ hold. We start proving the consistency of the IOLS estimator. This proof follows analogous steps as in Theorems 1 and 4 in Dominitz and Sherman (2005). From them, if $N(\phi)$ is an

ACM on $(\mathbb{B}, E_2)$, then $N(\phi)$ is also a contraction map. We prove item (i.) in Theorem 1 by using the standard fixed-point theorem. Subject to the conditions stated in Lemma 1, $N(\beta^j)$ is an ACM implying that $\left|N\left(\beta^{j-1}\right) - N(\beta)\right| \leq \kappa \left|\beta^{j-1} - \beta\right|$ holds. Identification on the population mapping gives $N(\beta) = \beta$. Lemma 7 yields that the sample mapping is also an ACM on $(\mathbb{B}, E_2)$ with a fixed point $\widehat{\beta}$ in the closed set $\mathbb{B}$, such that $\widehat{\beta} \in \mathbb{B}$. Bound the absolute difference between the fixed point from the sample mapping and β as

$$\left|\widehat{\beta} - \beta\right| \leq \left|\beta^j - \beta\right| + \left|\widehat{\beta} - \beta^j\right|. \tag{21}$$

To show that $\left|\beta^j - \beta\right|$ converges to zero, rewrite it as

$$\left|\beta^j - \beta\right| = \left|N\left(\beta^{j-1}\right) - N(\beta)\right| \leq \kappa \left|\beta^{j-1} - \beta\right|, \tag{22}$$

provided that $N(\phi)$ is an ACM and thus $N(\beta) = \beta$. Substituting recursively equation (22) yields $\left|\beta^j - \beta\right| \leq \kappa^j \left|\beta^0 - \beta\right|$. As $j \longrightarrow \infty$, $\left|\beta^j - \beta\right| = o(1)$, and hence the first term on the right-hand side of (21) converges to zero. It remains to show that $\left|\widehat{\beta} - \beta^j\right|$ has order $o_p(1)$. We bound $\left|\widehat{\beta} - \beta^j\right|$ using the auxiliary result (S.83) in Lemma 7, so that

$$\left|\widehat{\beta} - \beta^j\right| \leq \kappa^j \left|\beta^0 - \widehat{\beta}\right| + \left(\sum_{i=0}^{j-1}\right) \kappa^i \left[\sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right|\right]. \tag{23}$$

As $j \longrightarrow \infty$, with $\kappa \in (0,1]$, $\beta^0 \in \mathbb{B}$, $\widehat{\beta} \in \mathbb{B}$, and $\mathbb{B}$ is a closed ball centered at β, (23) reduces to

$$\left|\widehat{\beta} - \beta^j\right| \leq \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| \left[\frac{1}{1-\kappa}\right]. \tag{24}$$

Because the second term in brackets on the right-hand side of (24) is bounded and Lemma 5 yields that the first term has order $o_p(1)$, we have that the fixed point from the sample mapping is a consistent estimate of β, provided that $j \longrightarrow \infty$ and $T \longrightarrow \infty$.

We now turn our attention to the asymptotic distribution of the IOLS estimator. Rewrite $\sqrt{T}\left[\widehat{\beta} - \beta\right]$ as

$$\sqrt{T}\left[\widehat{\beta} - \beta\right] = \sqrt{T}\left[\widehat{N}_T(\beta) - \breve{N}_T(\beta)\right] + \sqrt{T}\left[\widehat{N}_T\left(\widehat{\beta}\right) - \widehat{N}_T(\beta)\right] + \sqrt{T}\left[\breve{N}_T(\beta) - \beta\right]. \tag{25}$$

Lemma 4 yields $\sqrt{T}\left[\widehat{N}_T(\beta)-\breve{N}_T(\beta)\right]=O_p\left(T^{-1/2}\right)$. Using the mean value theorem, the second term on the right-hand side of (25) can be rewritten as

$$\sqrt{T}\left[\widehat{N}_T\left(\widehat{\beta}\right)-\widehat{N}_T(\beta)\right]=\sqrt{T}\Big\{\left[\widehat{N}_T\left(\widehat{\beta}\right)-\breve{N}_T\left(\widehat{\beta}\right)\right]+\left[\breve{N}_T(\beta)-\widehat{N}_T(\beta)\right]\Big\}+ \sqrt{T}\Big\{\breve{\Lambda}_T\left(\widehat{\beta},\beta\right)\left[\widehat{\beta}-\beta\right]\Big\}, \quad (26)$$

where $\breve{\Lambda}_T\left(\widehat{\beta},\beta\right)=\int_0^1\breve{V}_T\left(\widehat{\beta}+\xi\left(\widehat{\beta}-\beta\right)\right)d\xi$. The first term on the right-hand side of (26) is $O_p\left(T^{-1/2}\right)$ following Lemma 4. Auxiliary result in the proof of Lemma 6 shows that $\breve{\Lambda}_T\left(\widehat{\beta},\beta\right)\left[\widehat{\beta}-\beta\right]$ converges uniformly to its population counterpart and (25) is given by

$$\sqrt{T}\left[\widehat{\beta}-\beta\right]=\sqrt{T}\left[\left[I_2-\Lambda\left(\widehat{\beta},\beta\right)\right]^{-1}\left[\breve{N}_T(\beta)-\beta\right]\right]. \quad (27)$$

From item (i.) in Theorem 1, $\widehat{\beta}$ converges in probability to β as $j\longrightarrow\infty$ with $T\longrightarrow\infty$, implying that $\Lambda\left(\widehat{\beta},\beta\right)$ converges in probability to $V(\beta)$. It follows that (27) reduces to:

$$\sqrt{T}\left[\widehat{\beta}-\beta\right]=\sqrt{T}\left[\left[I_2-V(\beta)\right]^{-1}\left[\breve{N}_T(\beta)-\beta\right]\right]. \quad (28)$$

Hence, as $T\longrightarrow\infty$ it remains to study the asymptotic distribution of $\sqrt{T}\left[\breve{N}_T(\beta)-\beta\right]$. Define $x_t=(y_{t-1},u_{t-1})'$, and recall that $\breve{N}_T(\beta)$ is evaluated at the true vector of parameters β. It follows that, as $T\longrightarrow\infty$, $\sqrt{T}\left[\breve{N}_T(\beta)-\beta\right]$ is given by

$$\sqrt{T}\left[\breve{N}_T(\beta)-\beta\right]=\sqrt{T}\left[\left[\sum_{t=2}^{T}x_tx_t'\right]^{-1}\sum_{t=2}^{T}x_tu_t\right]=\left[\frac{\sum_{t=2}^{T}x_tx_t'}{T}\right]^{-1}\left[\frac{1}{\sqrt{T}}\sum_{t=2}^{T}x_tu_t\right]. \quad (29)$$

Applying the central limit theorem for *mds* and because Assumptions A.1 and A.2 assure $\text{plim}\left[\frac{1}{T}\sum_{t=2}^{T}x_tx_t'\right]^{-1}=H^{-1}$, it follows that

$$\sqrt{T}\left[\breve{N}_T(\beta)-\beta\right]\xrightarrow{d}\mathcal{N}\left(0,\sigma_u^2H^{-1}\right). \quad (30)$$

Define $J:=\left[I_2-V(\beta)\right]^{-1}$ in (28) and using the result in (30), it follows that

$$\sqrt{T}\left[\widehat{\beta}-\beta\right]\xrightarrow{d}\mathcal{N}\left(0,\sigma_u^2JH^{-1}J'\right), \quad (31)$$

with

$$\Sigma_\beta := \sigma_u^2 J H^{-1} J' = \begin{pmatrix} -\frac{\left(-1+\beta_1^2\right)\left(1+2\beta_1\beta_2+\beta_2^2\right)}{\left(\beta_1+\beta_2\right)^2} & \frac{\left(-1+\beta_1^2\right)\left(1+\beta_1\beta_2\right)}{\left(\beta_1+\beta_2\right)^2} \\ \frac{\left(-1+\beta_1^2\right)\left(1+\beta_1\beta_2\right)}{\left(\beta_1+\beta_2\right)^2} & \frac{\left(1+\beta_1\beta_2\right)^2}{\left(\beta_1+\beta_2\right)^2} \end{pmatrix}. \tag{32}$$

Proof of Theorem 2: The proof of item i. follows from item i. in Theorem 1. To show item ii., we rewrite $\sqrt{T}\left[\widehat{\beta}-\beta\right]$ using the same arguments as in Theorem 1, such that

$$\sqrt{T}\left[\widehat{\beta}-\beta\right] = \sqrt{T}\left[\left[I_n - V\left(\beta\right)\right]^{-1}\left[\breve{N}_T\left(\beta\right)-\beta\right]\right]. \tag{33}$$

Because $\breve{N}_T\left(\beta\right)$ depends on the true vector of parameters β and $T \longrightarrow \infty$, rewrite the last term in (33) as:

$$\sqrt{T}\left[\breve{N}_T\left(\beta\right)-\beta\right] = \left[R'\left[\left(\frac{1}{T}\sum_{t=\bar{q}+1}^{T} X_t X_t'\right)\otimes I_K\right]R\right]^{-1}\left[\frac{1}{\sqrt{T}}\sum_{t=\bar{q}+1}^{T} R'\left(X_t \otimes I_K\right)U_t\right]. \tag{34}$$

Provided that $T \longrightarrow \infty$ and the fact that Assumptions B.1, B.2 and B.3 guarantee that the probability limit of the second moment matrices exist, the central limit theorem for *mds* can be used to show that (34) converges in distribution to

$$\sqrt{T}\left[\breve{N}_T\left(\beta\right)-\beta\right] \xrightarrow{d} \mathcal{N}\left(0, ZR'\left(H\otimes\Sigma_u\right)RZ'\right), \tag{35}$$

where $Z = \left[R'\left(H\otimes I_K\right)R\right]^{-1}$ and $H = \text{plim}\,\frac{1}{T}\sum_{t=\bar{q}+1}^{T} X_t X_t'$. Define $J := \left[I_n - V\left(\beta\right)\right]^{-1}$ and combine (33) with (35) to obtain the asymptotic distribution of the IOLS estimator for the general VARMA(p,q),

$$\sqrt{T}\left[\widehat{\beta}-\beta\right] \xrightarrow{d} \mathcal{N}\left(0, JZR'\left(H\otimes\Sigma_u\right)RZ'J'\right). \tag{36}$$

Table 1: Monte Carlo - Weak VARMA(1,1) models: Small Sized Systems, $K = 3$.

	DGP I						DGP II						DGP III					
	T=50, $n = 18$						T=50, $n = 18$						T=50, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	0.87	**0.82**	1.09	0.95	0.91	1.00	**0.94**	0.96	1.10	0.97	1.08	1.00	0.66	0.77	0.78	0.98	**0.61**
Share (%)	0%	**56%**	28%	0%	11%	6%	0%	**39%**	6%	0%	33%	22%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	97%	44%	73%	84%	85%	45%	98%	23%	84%	74%	91%	24%	99%	66%	96%	88%	95%	72%
	T=100, $n = 18$						T=100, $n = 18$						T=100, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	0.97	**0.91**	1.05	0.99	0.97	1.00	**0.98**	0.99	1.10	1.01	1.13	1.00	0.79	0.85	0.82	1.11	**0.75**
Share (%)	0%	**33%**	28%	6%	6%	28%	0%	**56%**	6%	6%	33%	0%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	99%	65%	89%	95%	94%	86%	99%	35%	94%	79%	97%	59%	99%	70%	99%	91%	100%	90%
	T=150, $n = 18$						T=150, $n = 18$						T=150, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.04	1.07	1.06	**0.99**	1.00	1.00	**0.95**	1.03	1.06	1.01	1.09	1.00	0.86	1.29	0.87	1.17	**0.82**
Share (%)	6%	**33%**	6%	6%	22%	28%	0%	**44%**	33%	11%	0%	11%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	100%	74%	81%	98%	97%	94%	100%	40%	82%	84%	99%	70%	100%	72%	98%	93%	100%	95%
	T=200, $n = 18$						T=200, $n = 18$						T=200, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.05	1.04	1.06	**1.00**	1.01	1.00	**0.94**	1.03	1.04	1.01	1.05	1.00	0.88	1.29	0.88	1.16	**0.85**
Share (%)	11%	28%	11%	6%	6%	**39%**	0%	**39%**	33%	11%	6%	11%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	100%	78%	88%	99%	98%	95%	100%	43%	86%	86%	99%	74%	100%	71%	100%	93%	100%	97%
	T=400, $n = 18$						T=400, $n = 18$						T=400, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.05	1.04	1.03	1.00	**1.00**	1.00	**0.88**	1.05	0.95	1.01	0.92	1.00	**0.90**	1.39	0.92	1.16	0.91
Share (%)	11%	28%	6%	6%	11%	**39%**	0%	**44%**	33%	6%	0%	17%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	100%	88%	95%	100%	99%	97%	100%	54%	89%	91%	100%	80%	100%	71%	100%	95%	100%	99%

We report results from weak VARMA(1,1) models simulated with different Kronecker indices. The first set of results reports results from DGP I, while the second and third set of results display results from weak VARMA(1,1) models simulated from DGPs II and III, respectively. Recall that DGP I and II set all the Kronecker indices to one, $\mathbf{p} = (1, 1, 1)'$, while DGP III sets $\mathbf{p} = (1, 0, 0)'$. DGPs I and II differ with respect to the eigenvalues driving the AR and MA parameter matrices. DGP I has all the eigenvalues associated with both the AR and MA parameter matrices set to 0.5, while DGP II has eigenvalues associated with the AR and MA components given by $(0.80, 0.20, 0.05)'$ and $(0.90, -0.02, -0.20)'$, respectively. The online Supplement gives more details on the different DGPs and restrictions imposed by the Echelon form transformation. The true vector of parameters in DGP III contains the estimates obtained by fitting a VARMA(1,1) model on the first rolling window of a dataset comprising only the three key macroeconomic variables studied in Section 5. Table S.1 displays the true values used to simulate DGPs I, II and III. Weak innovations are generated using the procedure described in Romano and Thombs (1996). n accounts for the number of free parameters in the model. MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share % is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage of Hannan and Kavalieris (1984); KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012); and MLE accounts for the maximum-likelihood estimator. The number of replications is set to 5000.

Table 2: Monte Carlo - Weak VARMA(1,1) models: Medium and Large Sized Systems, $K = 10$, $K = 20$, $K = 40$ and $K = 52$, with $T = 400$.

	DGP III					DGP IV					DGP V				
	K=10, $n = 20$					K=10, $n = 40$					K=10, $n = 60$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.99	2.00	**0.76**	1.17	1.00	0.94	1.57	**0.89**	1.22	1.00	0.97	1.46	**0.94**	1.20
Share (%)	0%	25%	5%	**70%**	0%	10%	23%	8%	**60%**	0%	17%	18%	17%	**48%**	0%
Convergence(%)	100%	98%	100%	100%	98%	100%	92%	99%	100%	95%	100%	83%	99%	100%	95%
	K=20, $n = 40$					K=20, $n = 80$					K=20, $n = 120$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	1.01	2.23	**0.79**	1.49	1.00	0.98	1.76	**0.87**	1.41	1.00	0.98	1.65	**0.88**	1.45
Share (%)	30%	15%	0%	**55%**	0%	19%	19%	1%	**61%**	0%	13%	14%	7%	**65%**	2%
Convergence(%)	100%	100%	100%	100%	82%	100%	97%	100%	100%	80%	100%	94%	100%	100%	75%
	K=40, $n = 80$					K=40, $n = 160$					K=40, $n = 240$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	**0.35**	1.30	0.46	1.07	1.00	**0.44**	1.30	0.59	1.28	1.00	**0.47**	1.36	0.62	1.37
Share (%)	4%	**56%**	0%	40%	0%	9%	**75%**	0%	16%	0%	10%	**83%**	0%	8%	0%
Convergence(%)	100%	100%	87%	100%	65%	100%	99%	87%	100%	62%	100%	100%	80%	99%	45%
	K=52, $n = 104$					K=52, $n = 208$					K=52, $n = 312$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	**0.41**	2.01	0.45	2.39	1.00	**0.43**	1.71	0.53	2.16	1.00	**0.48**	1.92	0.58	2.76
Share (%)	4%	**51%**	0%	45%	0%	4%	**67%**	0%	29%	0%	6%	**82%**	0%	12%	0%
Convergence(%)	100%	100%	81%	99%	36%	100%	100%	75%	97%	22%	100%	100%	63%	96%	8%

We report results from weak VARMA(1,1) models simulated with different Kronecker indices, system sizes and fixed sample size set as $T = 400$. The first set of results reports results from DGP III, while the second and third set of results display results from weak VARMA(1,1) models simulated from DGPs IV and V, respectively. Recall that DGP III sets the first Kronecker index to one and all the remaining indices to zero, $\mathbf{p} = (1, 0, 0, 0, ..., 0)'$; DGP IV sets the first two Kronecker indices to one and the remaining $K - 2$ indices to zero, $\mathbf{p} = (1, 1, 0, 0, ..., 0)'$; and DGP V sets the first three Kronecker indices to one and the remaining $K - 3$ indices to zero, $\mathbf{p} = (1, 1, 1, 0, ..., 0)'$. The online Supplement gives more details on the different DGPs and restrictions imposed by the Echelon form transformation. The true vectors of parameters in DGP III, IV and V contain the estimates obtained by fitting VARMA(1,1) models on the first rolling window of Dataset 1 in their respective system dimensions. Tables S.2 and S.3 display the true values used to simulate DGPs III, IV and V. Weak innovations are generated using the procedure described in Romano and Thombs (1996). n accounts for the number of free parameters in the model. MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share % is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage of Hannan and Kavalieris (1984); and KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012). The number of replications is set to 1000.

Table 3: Forecast Summary: VARMA Out-of-Sample Performance Relative to Alternative Group of Models

$K = 3$

	VARMA				$\mathbf{p}_{100}$				$\mathbf{p}_{110}$				$\mathbf{p}_{111}$				$\mathbf{p}_{HK}$			
	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR
Hor: 1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	67%	100%	100%	100%	33%
Hor: 2	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	67%	100%	100%	100%	67%
Hor: 3	100%	100%	100%	67%	67%	67%	100%	67%	67%	67%	100%	67%	100%	100%	100%	67%	100%	67%	100%	67%
Hor: 6	100%	100%	100%	100%	100%	67%	100%	100%	100%	100%	100%	100%	100%	67%	100%	100%	100%	67%	100%	100%
Hor: 9	67%	100%	0%	67%	67%	100%	0%	67%	67%	100%	0%	67%	67%	67%	0%	67%	67%	67%	0%	67%
Hor: 12	100%	67%	67%	100%	67%	33%	33%	33%	67%	33%	33%	33%	100%	33%	33%	67%	100%	67%	67%	100%

$K = 10$

	VARMA				$\mathbf{p}_{100}$				$\mathbf{p}_{110}$				$\mathbf{p}_{111}$				$\mathbf{p}_{HK}$			
	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR
Hor: 1	83%	100%	58%	42%	83%	100%	58%	42%	58%	100%	33%	17%	42%	92%	33%	0%	58%	100%	33%	25%
Hor: 2	100%	100%	92%	92%	92%	100%	83%	92%	100%	100%	92%	92%	50%	83%	50%	33%	67%	92%	58%	33%
Hor: 3	100%	100%	92%	100%	83%	83%	75%	58%	92%	83%	67%	67%	67%	75%	58%	58%	33%	58%	42%	8%
Hor: 6	100%	83%	100%	58%	100%	67%	100%	50%	100%	67%	92%	42%	67%	50%	67%	25%	83%	67%	83%	25%
Hor: 9	100%	58%	33%	75%	100%	50%	17%	50%	100%	42%	17%	50%	33%	25%	0%	33%	58%	25%	33%	25%
Hor: 12	92%	67%	67%	83%	58%	50%	58%	75%	58%	50%	58%	75%	67%	58%	33%	67%	58%	58%	42%	75%

$K = 20$

	VARMA				$\mathbf{p}_{100}$				$\mathbf{p}_{110}$				$\mathbf{p}_{111}$				$\mathbf{p}_{HK}$			
	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR
Hor: 1	67%	92%	67%	67%	67%	92%	58%	67%	58%	67%	33%	17%	42%	58%	17%	8%	67%	92%	50%	58%
Hor: 2	92%	100%	75%	92%	75%	92%	67%	83%	83%	92%	75%	83%	42%	67%	42%	42%	75%	92%	67%	83%
Hor: 3	92%	92%	67%	67%	83%	58%	17%	50%	83%	58%	33%	50%	75%	75%	50%	58%	50%	50%	17%	33%
Hor: 6	100%	67%	100%	58%	100%	58%	100%	50%	100%	50%	100%	33%	92%	42%	92%	17%	67%	50%	75%	17%
Hor: 9	100%	75%	33%	83%	100%	67%	33%	58%	100%	67%	33%	50%	50%	42%	8%	33%	50%	50%	17%	42%
Hor: 12	100%	75%	42%	67%	58%	58%	42%	42%	58%	58%	42%	50%	67%	75%	42%	50%	42%	42%	33%	33%

$K = 40$

	VARMA				$\mathbf{p}_{100}$				$\mathbf{p}_{110}$				$\mathbf{p}_{111}$				$\mathbf{p}_{HK}$			
	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR
Hor: 1	67%	100%	42%	83%	67%	100%	33%	67%	67%	100%	33%	33%	33%	100%	0%	25%	33%	75%	25%	42%
Hor: 2	75%	100%	42%	83%	67%	100%	33%	75%	67%	100%	33%	83%	33%	100%	0%	50%	50%	100%	33%	42%
Hor: 3	100%	100%	67%	83%	42%	100%	17%	42%	50%	100%	8%	50%	75%	100%	42%	75%	33%	100%	33%	25%
Hor: 6	100%	100%	42%	50%	100%	100%	42%	42%	92%	100%	42%	33%	75%	100%	42%	25%	75%	100%	25%	42%
Hor: 9	100%	100%	17%	50%	100%	100%	8%	33%	100%	100%	0%	17%	92%	100%	0%	17%	83%	100%	8%	33%
Hor: 12	67%	100%	42%	42%	33%	100%	42%	42%	33%	100%	42%	42%	33%	100%	42%	42%	50%	100%	42%	42%

$K = 52$

	VARMA				$\mathbf{p}_{100}$				$\mathbf{p}_{110}$				$\mathbf{p}_{111}$				$\mathbf{p}_{HK}$			
	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR	AR	VAR	FM	BVAR
Hor: 1	67%	100%	33%	67%	67%	100%	33%	67%	67%	100%	33%	33%	33%	100%	0%	0%	0%	67%	0%	33%
Hor: 2	67%	100%	33%	67%	67%	100%	33%	67%	67%	100%	33%	67%	33%	100%	0%	33%	33%	100%	33%	33%
Hor: 3	100%	100%	33%	100%	67%	100%	0%	67%	67%	100%	0%	67%	100%	100%	33%	100%	0%	100%	0%	0%
Hor: 6	100%	100%	67%	67%	100%	100%	67%	33%	100%	100%	67%	33%	67%	100%	67%	67%	67%	100%	67%	0%
Hor: 9	100%	100%	33%	67%	100%	100%	0%	33%	100%	100%	0%	33%	67%	100%	0%	33%	67%	100%	33%	67%
Hor: 12	67%	100%	33%	67%	33%	100%	33%	33%	67%	100%	33%	33%	33%	100%	33%	33%	33%	100%	33%	67%

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. For each system size, the first panel reports the frequency (in percentage points) for which at least one of the VARMA specifications ($\mathbf{p}_{100}$, $\mathbf{p}_{110}$, $\mathbf{p}_{111}$ and $\mathbf{p}_{HK}$) outperforms (delivers the lowest RelMSFE measures) the assigned group of competitors in a given forecast horizon. The second, third, fourth and fifth panels report the frequency the $\mathbf{p}_{100}$, $\mathbf{p}_{110}$, $\mathbf{p}_{111}$ and $\mathbf{p}_{HK}$ specifications, respectively, outperform the assigned group of competitors. We consider four groups of competitors. AR collects the AR(1) model; VAR contains the VAR(p^*) model, where p^* is obtained by minimizing the AIC criterion; FM gathers the factor model specifications, namely FM_{IC_3} and FM_{SC}; and BVAR aggregates the three Bayesian VAR models with the normal inverted Wishart prior which reproduces the principles of the Minnesota-type prior: BVAR_{SC}, $\text{BVAR}_{0.2}$ and BVAR_{opt}. Percentages are also computed across the four different datasets discussed in page 22 and specified in details in Table S.9. Values greater or equal than 50% are highlighted in bold.

Table 4: Forecast: IOLS Out-of-Sample Performance Relative to Alternative VARMA Estimators

K=3

	$\mathbf{p}_{100}$			$\mathbf{p}_{110}$			$\mathbf{p}_{111}$			$\mathbf{p}_{HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	**100%**	33%	**100%**	**100%**	**67%**	**67%**	**100%**	**100%**	33%	**100%**	**100%**	**100%**
Hor: 2	**100%**	**67%**	**100%**	**100%**	**67%**	**67%**	**100%**	**100%**	33%	**100%**	**100%**	**67%**
Hor: 3	**100%**	**67%**	**67%**	**100%**	**67%**	**67%**	**100%**	**67%**	**100%**	**100%**	**67%**	**67%**
Hor: 6	**100%**	**100%**	**100%**	**100%**	0%	**100%**	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**
Hor: 9	**100%**	**100%**	**67%**	**100%**	33%	**67%**	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**
Hor: 12	**100%**	**100%**	33%	**100%**	**100%**	**67%**	**100%**	**67%**	**67%**	**100%**	**67%**	**67%**

K=10

	$\mathbf{p}_{100}$			$\mathbf{p}_{110}$			$\mathbf{p}_{111}$			$\mathbf{p}_{HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	**75%**	**58%**	42%	42%	**58%**	**50%**	**58%**	42%	**50%**	25%	**67%**	17%
Hor: 2	**100%**	**75%**	**92%**	**92%**	**83%**	**92%**	42%	33%	**50%**	25%	25%	33%
Hor: 3	**100%**	**83%**	**83%**	**100%**	**67%**	**83%**	**92%**	**58%**	**67%**	**58%**	17%	17%
Hor: 6	**100%**	**75%**	**100%**	**100%**	**58%**	**75%**	**67%**	33%	**67%**	**67%**	42%	33%
Hor: 9	**100%**	**83%**	**100%**	**100%**	**75%**	**100%**	42%	**58%**	42%	**58%**	**50%**	33%
Hor: 12	**75%**	**92%**	**67%**	**92%**	**92%**	**75%**	**75%**	**75%**	**75%**	**92%**	**50%**	**58%**

K=20

	$\mathbf{p}_{100}$			$\mathbf{p}_{110}$			$\mathbf{p}_{111}$			$\mathbf{p}_{HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	**50%**	0%	42%	33%	**50%**	**58%**	**67%**	**58%**	**75%**	33%	**50%**	17%
Hor: 2	**83%**	17%	42%	**75%**	**50%**	**75%**	42%	25%	**75%**	**50%**	42%	0%
Hor: 3	**100%**	42%	33%	**100%**	**58%**	42%	**58%**	**50%**	**67%**	42%	42%	**50%**
Hor: 6	**100%**	42%	**75%**	**100%**	33%	**75%**	**100%**	25%	**58%**	**50%**	25%	17%
Hor: 9	**100%**	**92%**	**92%**	**100%**	**92%**	**92%**	**75%**	**58%**	**58%**	**67%**	42%	33%
Hor: 12	**83%**	**100%**	**75%**	**83%**	**92%**	**83%**	**92%**	**50%**	**75%**	**75%**	**92%**	33%

K=40

	$\mathbf{p}_{100}$			$\mathbf{p}_{110}$			$\mathbf{p}_{111}$			$\mathbf{p}_{HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	**58%**	**75%**	**92%**	**100%**	**92%**	**100%**	**100%**	**100%**	**100%**	42%	33%	**92%**
Hor: 2	**67%**	**100%**	**100%**	**92%**	**100%**	**100%**	**92%**	**92%**	**100%**	**50%**	**75%**	**100%**
Hor: 3	**92%**	**100%**	**92%**	**100%**	**100%**	**100%**	**92%**	**100%**	**100%**	42%	**92%**	**100%**
Hor: 6	**100%**	**75%**	**92%**	**100%**	**92%**	**75%**	**83%**	**92%**	**75%**	**67%**	**92%**	**83%**
Hor: 9	**100%**	**75%**	**92%**	**100%**	**75%**	**67%**	**83%**	**83%**	**100%**	**75%**	**75%**	**92%**
Hor: 12	**100%**	**100%**	**58%**	**100%**	**83%**	42%	**100%**	**100%**	**58%**	**58%**	**67%**	**75%**

K=52

	$\mathbf{p}_{100}$			$\mathbf{p}_{110}$			$\mathbf{p}_{111}$			$\mathbf{p}_{HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	0%	**67%**	**67%**
Hor: 2	**67%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	**100%**	**100%**	0%	**67%**	**67%**
Hor: 3	**67%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	33%	**67%**	**67%**
Hor: 6	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	**100%**	**100%**
Hor: 9	**100%**	**100%**	**100%**	**67%**	**100%**	**67%**	**67%**	**100%**	**100%**	**100%**	**100%**	**100%**
Hor: 12	33%	**67%**	**100%**	**67%**	**67%**	**100%**	33%	**100%**	**100%**	**67%**	33%	33%

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. VARMA models denoted by $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ have the Kronecker indices $\mathbf{p} = (p_1, p_2, ..., p_K)'$ set as $p_i = 1$ for $i \leq k$ and $p_i = 0$ for all $i > k$ with $k = 1$, $k = 2$ and $k = 3$, respectively. $\mathbf{p}_{HK}$ is the VARMA model with Kronecker indices obtained using the Hannan-Kavalieris procedure. For $K \leq 10$, we implement the Hannan-Kavalieris on every rolling window, while for $K > 10$ we implement the Hannan-Kavalieris procedure on the first rolling window and carry the optimal Kronecker indices to the remaining rolling windows. We report the frequency (in percentage points) which the IOLS estimator delivers lower RelMSFE measures than the assigned competitor. Percentages are also computed across the four different datasets discussed in page 22 and specified in details in Table S.9. DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage of Hannan and Kavalieris (1984) and KP is the three-step multivariate version of Koreisha and Pukkila (1990) as in Kascha (2012). Values greater or equal than 50% are highlighted in bold.

Supplement to "Estimation and Forecasting in Vector Autoregressive Moving Average Models for Rich Datasets"

Gustavo Fruet Dias*
Department of Economics and Business Economics, Aarhus University
CREATES

George Kapetanios
King's College London

June 7, 2016

Contents

*Corresponding author at: Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark. E-mail: gdias@econ.au.dk.

List of Figures

List of Tables

S. 1 Corollary 1

Proof of Corollary 1: We start by discussing how Lemmas 4, 5, 6, 7 and 8 hold when Assumption B.2 is replaced by Assumption B.2a. First, recall that Assumption B.2a implies that the disturbances are uncorrelated. It follows that the weak law of large numbers still applies, such that point-wise convergence of the infeasible sample mapping to its population counterpart (auxiliary result in Lemma 5) holds. Second, the unconditional moments of the weak VARMA(p,q) models are the same as their strong counterparts. Specifically, Assumptions B.1 and B.2a guarantee that the sample counterpart of the second moment matrices of the weak VARMA(p,q) models are $O_p(1)$ as $T \to \infty$. This is a key ingredient when proving Lemmas 4 and 5. Finally, Lemmas 6, 7 and 8 hold without changes following the validity of Lemmas 4, 5 and Assumptions B.1, B.2 and B.3.

To show consistency of the IOLS estimator, bound $\left|\widehat{\beta} - \beta\right|$ as in the proof of Theorem 1, such that

$$\left|\widehat{\beta} - \beta\right| \leq \left|\beta^j - \beta\right| + \left|\widehat{\beta} - \beta^j\right|. \tag{S.1}$$

Because the population mapping remains an ACM under Assumption B.2a, rewrite the first term on the right-hand side of (S.1) as

$$\left|\beta^j - \beta\right| \leq \kappa^j \left|\beta^0 - \beta\right|. \tag{S.2}$$

As $j \longrightarrow \infty$, it follows that $\kappa^j \left|\beta^0 - \beta\right|$ has order $o(1)$. It remains to study the properties of the second term on the right-hand side of (S.1). Because Lemmas 4-8 hold, bound $\left|\widehat{\beta} - \beta^j\right|$ as

$$\left|\widehat{\beta} - \beta^j\right| \leq \kappa^j \left|\beta^0 - \widehat{\beta}\right| + \left(\sum_{i=0}^{j-1}\right) \kappa^i \left[\sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right|\right]. \tag{S.3}$$

As $j \longrightarrow \infty$, with $\kappa \in (0, 1]$, $\left\{\beta^0, \widehat{\beta}\right\} \in \mathbb{B}$, and $\mathbb{B}$ is a closed ball centered at β, (S.3) reduces to

$$\left|\widehat{\beta} - \beta^j\right| \leq \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| \left[\frac{1}{1-\kappa}\right]. \tag{S.4}$$

Lemma 5 states that first term on the right-hand side of (S.4) has order $o_p(1)$, while $\kappa \in (0, 1]$

implies that $\frac{1}{1-\kappa}$ is $O(1)$. Combining these two results, $\left|\widehat{\beta}-\beta^j\right|$ is $o_p(1)$, which proves the first result of the Corollary.

Asymptotic normality is obtained in a similar manner as in Theorem 2. From (33), write $\sqrt{T}\left[\widehat{\beta}-\beta\right]$ as

$$\sqrt{T}\left[\widehat{\beta}-\beta\right]=\sqrt{T}\left[\left[I_n-V\left(\beta\right)\right]^{-1}\left[\breve{N}_T\left(\beta\right)-\beta\right]\right]. \tag{S.5}$$

The proof reduces to study the asymptotic distribution of $\sqrt{T}\left[\breve{N}_T\left(\beta\right)-\beta\right]$, so that as $T\to\infty$,

$$\begin{aligned}\sqrt{T}\left[\breve{N}_T\left(\beta\right)-\beta\right]=&\left[R'\left[\left(\frac{1}{T}\sum_{t=\bar{q}+1}^{T}X_tX_t'\right)\otimes I_K\right]R\right]^{-1}\times\\&\left[\frac{1}{\sqrt{T}}\sum_{t=\bar{q}+1}^{T}R'\left(X_t\otimes I_K\right)U_t\right].\end{aligned} \tag{S.6}$$

Recall that the last term is not a *mds*, because U_t is no longer an *iid* process. Assumption B.1 allows an VMA(∞) representation of Y_t of the form $Y_t=\sum_{i=0}^{\infty}\Theta_iU_{t-i}$ with $\Theta_0=I_K$. As discussed in Francq and Zakoian (1998), a stationary process which is a function of a finite number of current and lagged values of U_t satisfies a mixing property of the form Assumption B.2a. Using the VMA(∞) representation of Y_t, we partition X_t in (S.6) into $X_t=X_t^r+X_t^{r+}$, such that

$$X_t=\begin{pmatrix}Y_t-U_t\\Y_{t-1}\\\vdots\\Y_{t-p}\\U_{t-1}\\\vdots\\U_{t-q}\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{r}\Theta_iU_{t-i}\\\sum_{i=0}^{r}\Theta_iU_{t-1-i}\\\vdots\\\sum_{i=0}^{r}\Theta_iU_{t-p-i}\\U_{t-1}\\\vdots\\U_{t-q}\end{pmatrix}+\begin{pmatrix}\sum_{i=r+1}^{\infty}\Theta_iU_{t-i}\\\sum_{i=r+1}^{\infty}\Theta_iU_{t-1-i}\\\vdots\\\sum_{i=r+1}^{\infty}\Theta_iU_{t-p-i}\\0\\\vdots\\0\end{pmatrix}=X_t^r+X_t^{r+}. \tag{S.7}$$

Rewrite $\frac{1}{\sqrt{T}}\sum_{t=\bar{q}+1}^{T} R'\left(X_t \otimes I_K\right) U_t$ as

$$\frac{1}{\sqrt{T}}\sum_{t=\bar{q}+1}^{T} R'\left(X_t \otimes I_K\right) U_t = \frac{1}{\sqrt{T}}\sum_{t=\bar{q}+1}^{T} R'\left(X_t^r \otimes I_K\right) U_t + \frac{1}{\sqrt{T}}\sum_{t=\bar{q}+1}^{T} R'\left(X_t^{r+} \otimes I_K\right) U_t. \quad \text{(S.8)}$$

Auxiliary results in Dufour and Pelletier (2014, Theorem 4.2) show that the second term on the right-hand side of (S.8) converges uniformly to zero in T as $r \to \infty$. It follows that first term on the right-hand side of (S.8) satisfies the strong mixing conditions of the form Assumption B.2a. We are now in position to use the central limit theorem for strong mixing processes as in Ibragimov (1962) (see also Dufour and Pelletier (2014, Lemma A.2)). This yields $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} R'\left(X_t^r \otimes I_K\right) U_t \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_r\right)$. From Francq and Zakoian (1998), $\mathcal{I}_r \xrightarrow{p} \mathcal{I}$ as $r \to \infty$, such that as $T, r \to \infty$

$$\sqrt{T}\left[\check{N}_T\left(\beta\right) - \beta\right] \xrightarrow{d} \mathcal{N}\left(0, Z\mathcal{I}Z'\right), \quad \text{(S.9)}$$

where $\mathcal{I} = \sum_{\ell=-\infty}^{\infty} \mathbb{E}\left\{\left[R'\left(X_t \otimes I_K\right) U_t\right]\left[R'\left(X_{t-\ell} \otimes I_K\right) U_{t-\ell}\right]'\right\}$, and $Z = \left[R'\left(H \otimes I_K\right) R\right]^{-1}$ with $H = \operatorname{plim} \frac{1}{T}\sum_{t=\bar{q}+1}^{T} X_t X_t'$. The final result of this corollary is therefore obtained by combining the first element of the right-hand side of (S.5) with (S.9), such that

$$\sqrt{T}\left[\widehat{\beta} - \beta\right] \xrightarrow{d} \mathcal{N}\left(0, JZ\mathcal{I}Z'J'\right), \quad \text{(S.10)}$$

where $J := \left[I_n - V\left(\beta\right)\right]^{-1}$.

S. 2 Auxiliary Lemmas

S. 2.1 ARMA(1,1) model

Proof of Lemma 1: We mirror our proof in Lemma 5 in Dominitz and Sherman (2005). Fix $\phi, \gamma \in \mathbb{B}$, where $\mathbb{B}$ is the set of all possible parameter values satisfying Assumption A.1. There also exists a ϕ^* located in the segment line between ϕ and γ, such that $\left|N\left(\phi\right) - N\left(\gamma\right)\right| = \left|V\left(\phi^*\right)\left[\phi - \gamma\right]\right|$ holds. Using this result, we are in a position to define a bound that is function of the gradient of the population mapping evaluated on β, such that

$$\left|N\left(\phi\right) - N\left(\gamma\right)\right| = \left|V\left(\phi^*\right)\left[\phi - \gamma\right]\right| \leq \left|V\left(\beta\right)\left[\phi - \gamma\right]\right| + \left|\left[V\left(\phi^*\right) - V\left(\beta\right)\right]\left[\phi - \gamma\right]\right|. \quad \text{(S.11)}$$

From Dominitz and Sherman (2005), it suffices to show that the maximum eigenvalue of $V(\beta)$ is less than one in absolute value. To this purpose, define $V\left(\beta^j\right) = \frac{\partial N\left(\beta^{j\prime}\right)}{\partial \beta^{j\prime}}$ as the gradient from the population mapping on the $(j+1)^{th}$ iteration,

$$V\left(\beta^j\right) = \begin{bmatrix} \frac{\partial \beta_1^{j+1}}{\partial \beta_1^j} & \frac{\partial \beta_1^{j+1}}{\partial \beta_2^j} \\ \frac{\partial \beta_2^{j+1}}{\partial \beta_1^j} & \frac{\partial \beta_2^{j+1}}{\partial \beta_2^j} \end{bmatrix}. \tag{S.12}$$

Using the partitioned regression result, we obtain individual expressions for the OLS estimates of β obtained from the population mapping at each iteration:

$$\beta_1^{j+1} = \mathbb{E}\left[y_{t-1}^2\right]^{-1} \mathbb{E}\left[y_{t-1}\left[y_t - u_{t-1}^j \beta_2^{j+1}\right]\right], \tag{S.13}$$

$$\begin{aligned} \beta_2^{j+1} = & \left[\mathbb{E}\left[(u_{t-1}^j)^2\right] - \mathbb{E}\left[u_{t-1}^j y_{t-1}\right] \mathbb{E}\left[y_{t-1}^2\right]^{-1} \mathbb{E}\left[u_{t-1}^j y_{t-1}\right]\right]^{-1} \times \\ & \left[\mathbb{E}\left[u_{t-1}^j y_t\right] - \mathbb{E}\left[u_{t-1}^j y_{t-1}\right] \mathbb{E}\left[y_{t-1}^2\right]^{-1} \mathbb{E}\left[y_{t-1} y_t\right]\right]. \end{aligned} \tag{S.14}$$

Using the invertibility condition to express estimates of the lagged disturbances as an AR(∞), we have

$$\begin{aligned} \frac{\partial \beta_1^{j+1}}{\partial \beta_1^j} = & \left\{\mathbb{E}\left[y_{t-1}^2\right]^{-1} \mathbb{E}\left[y_{t-1}\left[\left(1+\beta_2^j L\right)^{-1} y_{t-2}\right]\right]\right\} \beta_2^{j+1} - \\ & \mathbb{E}\left[y_{t-1}^2\right]^{-1} \mathbb{E}\left[y_{t-1} u_{t-1}^j\right] \frac{\partial \beta_2^{j+1}}{\partial \beta_1^j}. \end{aligned} \tag{S.15}$$

Evaluating (S.15) on the true vector of parameters β, the first element of (S.12) reduces to

$$\left.\frac{\partial \beta_1^{j+1}}{\partial \beta_1^j}\right|_\beta = \left(\frac{1}{\sigma_y^2}\right)\left[\sum_{i=0}^{\infty} (-\beta_2)^i \rho_{1+i}\right] \beta_2 - \left(\frac{\sigma_u^2}{\sigma_y^2}\right)\left[\left.\frac{\partial \beta_2^{j+1}}{\partial \beta_1^j}\right|_\beta\right], \tag{S.16}$$

where $\mathbb{E}\left[y_{t-1}^2\right] = \sigma_y^2 = \frac{\left(1+\beta_2^2+2\beta_1\beta_2\right)\sigma_u^2}{\left(1-\beta_1^2\right)}$ is the variance of the ARMA(1,1) process, $\mathbb{E}\left[y_{t-1}u_{t-1}\right] = \sigma_u^2$ is the variance of the disturbances and $\mathbb{E}\left[y_t y_{t-\ell}\right] = \rho_\ell = \beta_1^{\ell-1}\left(\beta_1\sigma_y^2 + \beta_2\sigma_u^2\right)$ is the auto-covariance of lag ℓ.

Similarly to (S.15), the second element in the first row of (S.12) is

$$\begin{aligned} \frac{\partial \beta_1^{j+1}}{\partial \beta_2^j} = & \left\{\mathbb{E}\left[y_{t-1}^2\right]^{-1} \mathbb{E}\left[y_{t-1}\left[\left(1+\beta_2^j L\right)^{-1} u_{t-2}^j\right]\right]\right\} \beta_2^{j+1} - \\ & \mathbb{E}\left[y_{t-1}^2\right]^{-1} \mathbb{E}\left[y_{t-1} u_{t-1}^j\right] \frac{\partial \beta_2^{j+1}}{\partial \beta_2^j}. \end{aligned} \tag{S.17}$$

Evaluating (S.17) on the true vector of parameters β, the second element in the first row of (S.12) reduces to

$$\left.\frac{\partial\beta_1^{j+1}}{\partial\beta_2^j}\right|_\beta = \left(\frac{1}{\sigma_y^2}\right)\left[\sum_{i=0}^{\infty}(-\beta_2)^i\,\vartheta_{1+i}\right]\beta_2 - \left(\frac{\sigma_u^2}{\sigma_y^2}\right)\left[\left.\frac{\partial\beta_2^{j+1}}{\partial\beta_2^j}\right|_\beta\right], \tag{S.18}$$

with $\mathbb{E}\left[y_t u_{t-\ell}\right] = \vartheta_\ell = \beta_1^{\ell-1}\left[\sigma_u^2\left(\beta_1+\beta_2\right)\right]$. Computing the elements in the second row of (S.12) in a similar manner as in (S.16) and (S.18), we have

$$\begin{aligned}\left.\frac{\partial\beta_2^{j+1}}{\partial\beta_1^j}\right|_\beta = &-2\left[\sigma_u^2-\frac{\left(\sigma_u^2\right)^2}{\sigma_y^2}\right]^{-2}\left[\vartheta_1-\frac{\sigma_u^2\rho_1}{\sigma_y^2}\right]\left[\left(\frac{\sigma_u^2}{\sigma_y^2}\right)\left(\sum_{i=0}^{\infty}(-\beta_2)^i\,\rho_{1+i}\right)\right]+\\ &\left[\sigma_u^2-\frac{\left(\sigma_u^2\right)^2}{\sigma_y^2}\right]^{-1}\left[-\left(\sum_{i=0}^{\infty}(-\beta_2)^i\,\rho_{2+i}\right)+\left(\frac{\rho_1}{\sigma_y^2}\right)\left(\sum_{i=0}^{\infty}(-\beta_2)^i\,\rho_{1+i}\right)\right],\end{aligned} \tag{S.19}$$

$$\begin{aligned}\left.\frac{\partial\beta_2^{j+1}}{\partial\beta_2^j}\right|_\beta = &-2\left[\sigma_u^2-\frac{\left(\sigma_u^2\right)^2}{\sigma_y^2}\right]^{-2}\left[\vartheta_1-\frac{\sigma_u^2\rho_1}{\sigma_y^2}\right]\left[\left(\frac{\sigma_u^2}{\sigma_y^2}\right)\left(\sum_{i=0}^{\infty}(-\beta_2)^i\,\vartheta_{1+i}\right)\right]+\\ &\left[\sigma_u^2-\frac{\left(\sigma_u^2\right)^2}{\sigma_y^2}\right]^{-1}\left[-\left(\sum_{i=0}^{\infty}(-\beta_2)^i\,\vartheta_{2+i}\right)+\left(\frac{\rho_1}{\sigma_y^2}\right)\left(\sum_{i=0}^{\infty}(-\beta_2)^i\,\vartheta_{1+i}\right)\right].\end{aligned} \tag{S.20}$$

From (S.16), (S.18), (S.19) and (S.20) and using the fact that $\sum_{i=0}^{\infty}(-\beta_2)^i\,\rho_{1+i} = \frac{\beta_1\sigma_y^2+\beta_2\sigma_u^2}{1+\beta_1\beta_2}$, $\sum_{i=0}^{\infty}(-\beta_2)^i\,\rho_{2+i} = \frac{\beta_1\left(\beta_1\sigma_y^2+\beta_2\sigma_u^2\right)}{1+\beta_1\beta_2}$, $\sum_{i=0}^{\infty}(-\beta_2)^i\,\vartheta_{1+i} = \frac{\left(\beta_1+\beta_2\right)\sigma_u^2}{1+\beta_1\beta_2}$ and $\sum_{i=0}^{\infty}(-\beta_2)^i\,\vartheta_{2+i} = \frac{\left(\beta_1+\beta_2\right)\beta_1\sigma_u^2}{1+\beta_1\beta_2}$, we have that (S.12) evaluated at β, reduces to

$$V(\beta) = \left.\frac{\partial N\left(\beta^j\right)}{\partial\beta^{j\prime}}\right|_\beta = \begin{pmatrix}\frac{\beta_2}{\beta_1+\beta_2} & \frac{\beta_2\left(1-\beta_1^2\right)}{(\beta_1+\beta_2)(1+\beta_1\beta_2)}\\ \frac{-\beta_2}{\beta_1+\beta_2} & \frac{-\beta_2\left(1-\beta_1^2\right)}{(\beta_1+\beta_2)(1+\beta_1\beta_2)}\end{pmatrix}. \tag{S.21}$$

Note that (S.21) does not depend on σ_u^2, implying that Lemma 1 holds for any value assigned to the variance of the disturbances. The two eigenvalues associated with (S.21) are given by

$$\lambda_1 = 0, \tag{S.22}$$

$$\lambda_2 = \frac{\beta_1\beta_2}{(1+\beta_1\beta_2)}. \tag{S.23}$$

Because $\lambda_1 = 0$, we only need to show that $|\lambda_2| < 1$ to prove that the population mapping is an ACM. A sufficient condition is $\beta_1 \beta_2 > -1/2$.

Lemma 3 *Suppose Assumptions A.1 and A.2 hold. Then, as* $T \longrightarrow \infty$, $\breve{N}_T(\phi)$ *is stochastically equicontinuous.*

Proof of Lemma 3: We prove Lemma 3 by establishing the Lipschitz condition of $\breve{N}_T(\phi)$ similarly as in Lemma 2.9 in Newey and McFadden, 1994, pg. 2138. We need to show that $\sup_{\phi^* \in \mathbb{B}} \left\| \breve{V}_T(\phi^*) \right\| = O_p(1)$, where $\breve{V}_T(\phi^*)$ is the gradient of the infeasible sample mapping evaluated at any vector of estimates ϕ^*, such that $\phi^* \in \mathbb{B}$ satisfying Assumption A.1. Recall that the infeasible sample mapping for an univariate ARMA(1,1) reduces to

$$\breve{N}_T\left(\breve{\beta}^j\right) = \left[\left(\frac{1}{T-1}\sum_{t=2}^{T} x_{\infty,t}^{j} x_{\infty,t}^{j\prime}\right)^{-1}\right]\left[\frac{1}{T-1}\sum_{t=2}^{T} x_{\infty,t}^{j} y_t\right], \tag{S.24}$$

where $x_{\infty,t}^{j} = \left(y_{t-1}, u_{t-1}^{j}\right)'$. Bound the norm of the difference of the infeasible sample mapping evaluated at different points satisfying Assumption A.1 as

$$\sup_{\phi,\gamma \in \mathbb{B}} \left\| \breve{N}_T(\phi) - \breve{N}_T(\gamma) \right\| \leq \sup_{\phi^* \in \mathbb{B}} \left\| \breve{V}_T(\phi^*) \right\| \sup_{\phi,\gamma \in \mathbb{B}} \|\phi - \gamma\|, \tag{S.25}$$

where $\phi, \gamma, \phi^* \in \mathbb{B}$ and $\phi^* = (\phi_1^*, \phi_2^*)'$ lies on the segment line between ϕ and γ. The second step consists of computing the sample gradient, because $\sup_{\phi,\gamma \in \mathbb{B}} \|\phi - \gamma\|$ has order $O_p(1)$. Note that we need to define $\breve{V}_T(\phi^*)$ in a generic way such that it can be evaluated at any vector of estimates on any possible iteration. Using the same steps as in Lemma 1, the elements of $\breve{V}_T\left(\beta^j\right)$ evaluated at $\phi^* \in \mathbb{B}$ reduce to:

$$\left.\frac{\partial \breve{\beta}_1^{j+1}}{\partial \breve{\beta}_1^j}\right|_{\phi^*} = \left(\frac{1}{\breve{\sigma}_y^2}\right)\left[\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\rho}_{1+i}\right]\phi_2^* - \left(\frac{\breve{\zeta}_u^2}{\breve{\sigma}_y^2}\right)\left[\left.\frac{\partial \breve{\beta}_2^{j+1}}{\partial \breve{\beta}_1^j}\right|_{\phi^*}\right], \tag{S.26}$$

$$\left.\frac{\partial \breve{\beta}_1^{j+1}}{\partial \breve{\beta}_2^j}\right|_{\phi^*} = \left(\frac{1}{\breve{\sigma}_y^2}\right)\left[\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\delta}_{1+i}\right]\phi_2^* - \left(\frac{\breve{\delta}_0}{\breve{\sigma}_y^2}\right)\left[\left.\frac{\partial \breve{\beta}_2^{j+1}}{\partial \breve{\beta}_2^j}\right|_{\phi^*}\right], \tag{S.27}$$

$$\begin{aligned}\left.\frac{\partial \breve{\beta}_2^{j+1}}{\partial \breve{\beta}_1^j}\right|_{\phi^*} = -2&\left[\breve{\zeta}_u^2 - \frac{\left(\breve{\delta}_0^2\right)^2}{\breve{\sigma}_y^2}\right]^{-2}\left[\breve{\delta}_1 - \frac{\breve{\delta}_0^2\breve{\rho}_1}{\breve{\sigma}_y^2}\right]\times \\ &\left[-\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\xi}_{1+i} + \left(\frac{\breve{\delta}_0^2}{\breve{\sigma}_y^2}\right)\left(\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\rho}_{1+i}\right)\right] + \\ &\left[\breve{\zeta}_u^2 - \frac{\left(\breve{\delta}_0^2\right)^2}{\breve{\sigma}_y^2}\right]^{-1}\left[-\left(\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\rho}_{2+i}\right) + \left(\frac{\breve{\rho}_1}{\breve{\sigma}_y^2}\right)\left(\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\rho}_{1+i}\right)\right],\end{aligned} \tag{S.28}$$

$$\begin{aligned}\left.\frac{\partial \breve{\beta}_2^{j+1}}{\partial \breve{\beta}_2^j}\right|_{\phi^*} = -2&\left[\breve{\zeta}_u^2 - \frac{\left(\breve{\delta}_0^2\right)^2}{\breve{\sigma}_y^2}\right]^{-2}\left[\breve{\delta}_1 - \frac{\breve{\delta}_0^2\breve{\rho}_1}{\breve{\sigma}_y^2}\right]\left[\left(\frac{\breve{\delta}_0^2}{\breve{\sigma}_y^2}\right)\left(\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\delta}_{1+i}\right)\right] + \\ &\left[\breve{\zeta}_u^2 - \frac{\left(\breve{\delta}_0^2\right)^2}{\breve{\sigma}_y^2}\right]^{-1}\left[-\left(\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\delta}_{2+i}\right) + \left(\frac{\breve{\rho}_1}{\breve{\sigma}_y^2}\right)\left(\sum_{i=0}^{\infty}\left(-\phi_2^*\right)^i \breve{\delta}_{1+i}\right)\right],\end{aligned} \tag{S.29}$$

where $\breve{\zeta}_u^2 = \frac{1}{T-1}\sum_{t=2}^T u_t^j u_t^j$, $\breve{\delta}_0 = \frac{1}{T-1}\sum_{t=2}^T y_t u_t^j$, $\breve{\delta}_\ell = \frac{1}{T-1}\sum_{t=2}^T y_t u_{t-\ell}^j$, $\breve{\xi}_\ell = \frac{1}{T-1}\sum_{t=2}^T y_{t-\ell} u_t^j$, $\breve{\sigma}_y^2 = \frac{1}{T-1}\sum_{t=2}^T y_t^2$ and $\breve{\rho}_\ell = \frac{1}{T-1}\sum_{t=2}^T y_t y_{t-\ell}$. These quantities are all averages, and hence as $T \longrightarrow \infty$, they converge to their population counterparts. It is important to remark on two distinct results: first, we have that $\breve{\sigma}_y^2$ and $\breve{\rho}_\ell$ are quantities that do not depend on ϕ^*, implying that $\breve{\sigma}_y^2 \xrightarrow{p} \sigma_y^2$ and $\breve{\rho}_\ell \xrightarrow{p} \rho_\ell$ for all $\phi^* \in \mathbb{B}$ as $T \longrightarrow \infty$. These are the population moments generated by the ARMA(1,1) model and therefore depend only on β and σ_u^2. Second, we have that $\breve{\zeta}_u^2$, $\breve{\delta}_0$, $\breve{\delta}_\ell$ and $\breve{\xi}_\ell$ for $\ell \geq 1$ converge to finite quantities. Note that we do not require these quantities to converge to moments evaluated at the true vector of parameters β, but to some finite quantities that will depend on ϕ^*. Hence, considering some vector of estimates ϕ^*, we have that as $T \longrightarrow \infty$, the weak law of large numbers yields:

$$\breve{\delta}_0 \xrightarrow{p} \delta_0 = \frac{1}{(1+\beta_1\phi_2^*)}\left[\beta_1\left(\rho_1 - \phi_1^*\sigma_y^2\right)\sigma_u^2 + \beta_2\left(\vartheta_1 - \left(\phi_1^* + \phi_2^*\right)\sigma_u^2\right)\right], \tag{S.30}$$

$$\breve{\delta}_\ell \xrightarrow{p} \delta_\ell = \beta_1^{\ell-1}\left[\beta_1\delta_0 + \beta_2\sigma_u^2\right], \quad \ell \geq 1, \tag{S.31}$$

$$\breve{\zeta}_u^2 \xrightarrow{p} \zeta_u^2 = \frac{1}{(1-\phi_2^*)}\left[\left(1+\phi_1^{*^2}\right)\sigma_y^2 - 2\phi_1^*\rho_1 - 2\phi_2^*\delta_1 + 2\phi_1^*\phi_2^*\delta_0\right], \tag{S.32}$$

$$\breve{\xi}_1 \xrightarrow{p} \xi_1 = \rho_1 - \phi_1^*\sigma_y^2 - \phi_2^*\delta_0, \tag{S.33}$$

$$\begin{aligned}\breve{\xi}_\ell \xrightarrow{p} \xi_\ell = \rho_\ell + \left[\textstyle\sum_{i=2}^{\ell}(-1)^{\ell-2}\ (-1)^{\ell-1}\left(-\phi_2^*\right)^{\ell-i}\left(\phi_1^*+\phi_2^*\right)\rho_{i-1}\right] + \\ + (-1)^{\ell}\left[\left(-\phi_2^*\right)^{\ell-1}\phi_1^*\sigma_y^2 + \left(\phi_2^*\right)^{\ell}\delta_0\right], \quad \ell > 1.\end{aligned} \tag{S.34}$$

From Assumption A.1, we have that the $\sum_{i=0}^{\infty}|-\phi_2^*| < \infty$, $\sum_{i=0}^{\infty}|-\beta_2| < \infty$ and $\sum_{i=0}^{\infty}|-\beta_1| < \infty$ for all $\phi_2^*, \beta_1, \beta_2 \in \mathbb{B}$, implying that the infinite summations in (S.26), (S.27), (S.28) and (S.29) are summable. Following that, it is enough to show that $\left[\zeta_u^2 - \frac{(\delta_0^2)^2}{\sigma_y^2}\right]$ is bounded away from zero for all $\phi^*, \beta_1, \beta_2 \in \mathbb{B}$, to obtain $\sup_{\phi^*\in\mathbb{B}}\left\|\breve{V}_T(\phi^*)\right\| = O_p(1)$ as $T \longrightarrow \infty$. This is equivalent showing the parameters ϕ^*, β_1, β_2 that solve (S.35) lie outside $\mathbb{B}$:

$$\begin{aligned}\left[\zeta_u^2 - \frac{\left(\delta_0^2\right)^2}{\sigma_y^2}\right] = -\left(\phi_1^*+\phi_2^*\right)^2\left(-1+\beta_2\ \left(1-\beta_2+\beta_1\left(-1+\phi_2^*\right)+\phi_2^*\right)\right)\right] \times \\ \frac{\left[\left(1+\beta_2\left(1+\beta_1+\beta_2+\left(-1+\beta_1\right)\phi_2^*\right)\right)\left(\sigma_u^2\right)^2\right]}{\left[\left(-1+\beta_1^2\right)\left(1+\beta_1\phi_2^*\right)^2\left(-1+\phi_2^{*^2}\right)\right]} = 0.\end{aligned} \tag{S.35}$$

If this is the case, we have that $\left[\zeta_u^2 - \frac{(\delta_0^2)^2}{\sigma_y^2}\right]^{-1} = O_p(1)$ and $\left[\zeta_u^2 - \frac{(\delta_0^2)^2}{\sigma_y^2}\right]^{-2} = O_p(1)$ in (S.28) and (S.29) for all $\phi_2^*, \beta_1, \beta_2 \in \mathbb{B}$. By solving (S.35), we obtain multiple solutions that depend on the following four parameters: β_1, β_2, ϕ_1^* and ϕ_2^*, such that

$$\beta_1 = \left\{\frac{1-\beta_2+\beta_2^2-\beta_2\phi_2^*}{\beta_2\left(-1+\phi_2^{*^2}\right)}, \quad \frac{-1-\beta_2-\beta_2^2+\beta_2\phi_2^*}{\beta_2\left(1+\phi_2^{*^2}\right)}\right\}, \tag{S.36}$$

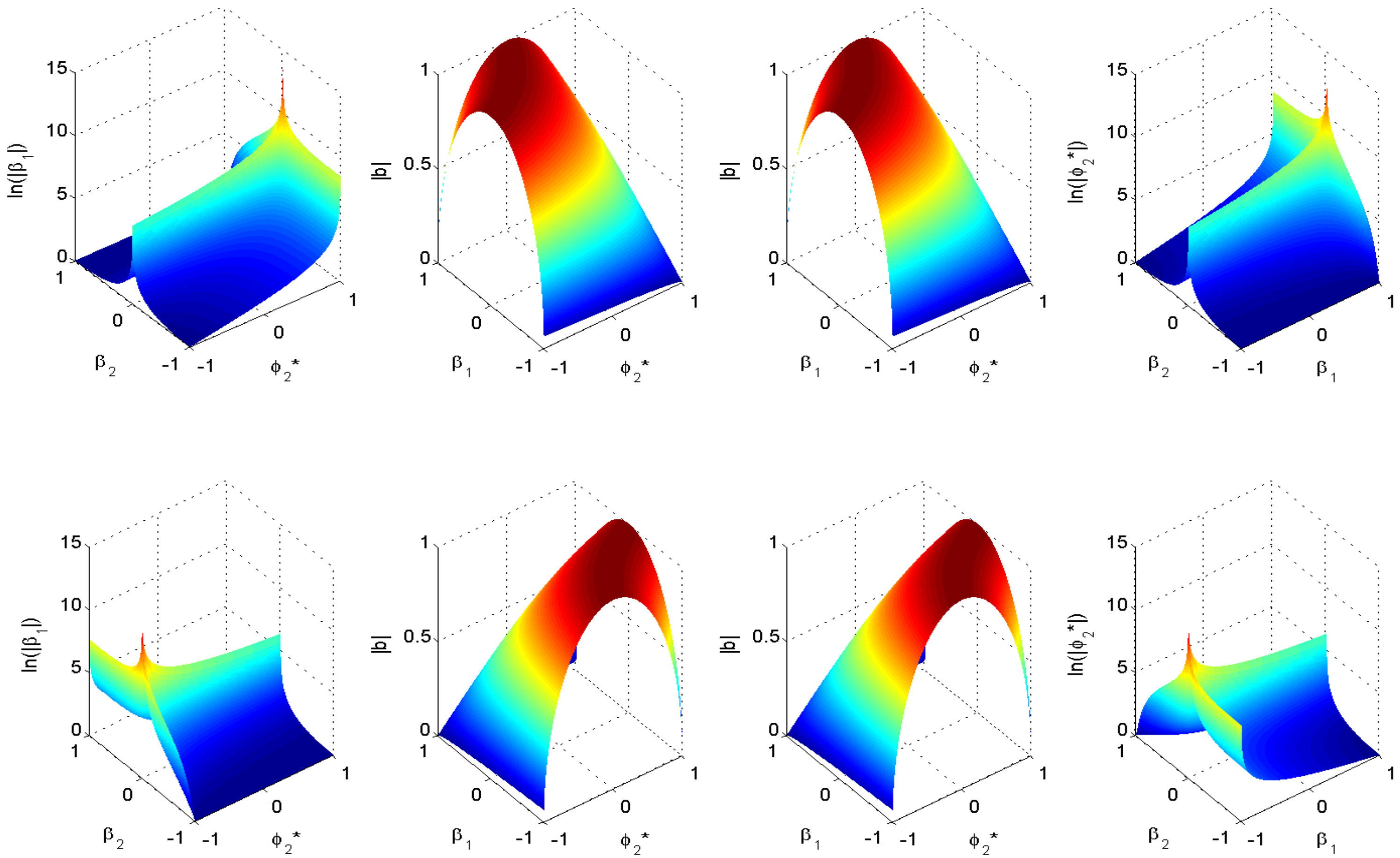
$$\begin{aligned}\beta_2 = \left\{\tfrac{1}{2}\left[1-\beta_1+\phi_2^*+\beta_1\phi_2^* \pm \sqrt{-4+\left(-1+\beta_1-\phi_2^*-\beta_1\phi_2^*\right)^2}\right],\right. \\ \left.\tfrac{1}{2}\left[-1-\beta_1+\phi_2^*-\beta_1\phi_2^* \pm \sqrt{-4+\left(1+\beta_1-\phi_2^*+\beta_1\phi_2^*\right)^2}\right]\right\},\end{aligned} \tag{S.37}$$

$$\phi_2^* = \left\{\frac{-1-\beta_2-\beta_2\beta_1-\beta_2^2}{(-1+\beta_1)\beta_2}, \quad \frac{1-\beta_2+\beta_2\beta_1+\beta_2^2}{(1+\beta_1)\beta_2}\right\}, \tag{S.38}$$

$$\phi_1^* = -\phi_2^*. \tag{S.39}$$

Solution (S.39) is ruled out by Assumption A.1. We tackle the remaining solutions through a numerical grid search. We show that the solutions given by (S.36), (S.37) and (S.38) lie outside $\mathbb{B}$ and thus violate Assumption A.1. We perform a numerical grid search on the $\phi_2^*, \beta_1, \beta_2 \in \mathbb{B}$ to obtain the solutions in (S.36), (S.37) and (S.38). The grid is fixed to 0.001 and the maximum and minimum values assigned to ϕ_2^*, β_1 and β_2 are 0.99 and -0.99, respectively. The first column in Figure S.1 displays the two solutions of (S.36). We show that $\ln|\beta_1|$ is bounded away from zero for all $\phi_2^*, \beta_2 \in \mathbb{B}$, implying that β_1 obtained from (S.36) is greater than one and therefore does not satisfy Assumption A.1. The second and third column in Figure S.1 bring the four solutions associated with (S.37). We find that the solutions are complex numbers for all $\phi_2^*, \beta_1 \in \mathbb{B}$. We illustrate this finding by representing the solutions in (S.37) as $\beta_2 = a + bi$ and plotting the absolute value of the imaginary part $|b|$. We show that $|b|$ is bounded away from zero, implying that β_2 in (S.37) is always a complex number and thus does not satisfy Assumption A.1. The third column in Figure S.1 presents graphs showing the two solutions of (S.38). We illustrate the results by plotting $\ln|\phi_2^*|$. We show that $\ln|\phi_2^*|$ is bounded away from zero for all $\beta_1, \beta_2 \in \mathbb{B}$. Hence, Figure S.1 shows that the solutions obtained in (S.36), (S.37) and (S.38) are not supported by Assumption A.1, implying that $\left[\zeta_u^2 - \frac{\left(\delta_0^2\right)^2}{\sigma_y^2}\right] \neq 0$ for all $\phi^*, \beta_1, \beta_2 \in \mathbb{B}$. It follows that $\sup_{\phi^* \in \mathbb{B}} \breve{V}_T(\phi^*) = O_p(1)$, yielding $\sup_{\phi^* \in \mathbb{B}} \left\| \breve{V}_T(\phi^*) \right\| = O_p(1)$, which proves Lemma 3.

Figure S.1: Numerical Solution, Lemma 3



We plot the solutions obtained from (S.36), (S.37) and (S.38) using a grid search of $\phi_2^*, \beta_1, \beta_2 \in \mathbb{B}$. The grid is fixed to 0.001 and the maximum and minimum values assigned to ϕ_2^*, β_1 and β_2 are 0.99 and -0.99, respectively. The first column plots $\ln|\beta_1|$ obtained through (S.36); the second and third columns plot $|b|$, the imaginary part of the complex solution of (S.38); and the fourth column displays $\ln|\phi_2^*|$ obtained using (S.38).

S. 2.2 VARMA(p,q) model

Proof of Lemma 2: Recall that the sample mapping in Definition 1 item (i.) is given by

$$\widehat{\beta}^{j+1} = \widehat{N}_T\left(\widehat{\beta}^j\right) = \left[\left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T} \widetilde{X}_t^{j\prime}\widetilde{X}_t^j\right)^{-1}\right]\left[\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T} \widetilde{X}_t^{j\prime} Y_t\right], \tag{S.40}$$

where $\widetilde{X}_t^j = \left[\left(\widehat{X}_t^{j\prime} \otimes I_K\right) R\right]$ has dimensions $(K \times n)$, $\bar{q} = \max\{p, q\}$, and n is the number of free parameters in the model. Using (S.40), we write the sample gradient evaluated at $\widehat{\beta}^j$ as

$$\widehat{V}_T\left(\widehat{\beta}^j\right) = \left\{\left[I_1 \otimes \left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T} \tilde{X}_t^{j\prime}\tilde{X}_t^j\right)^{-1}\right]\frac{\partial \text{vec}\left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\tilde{X}_t^{j\prime}Y_t\right)}{\partial\widehat{\beta}^{j\prime}} + \right.$$
$$\left.\left[\left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\tilde{X}_t^{j\prime}Y_t\right)' \otimes I_n\right]\frac{\partial \text{vec}\left(\left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\tilde{X}_t^{j\prime}\tilde{X}_t^j\right)^{-1}\right)}{\partial\widehat{\beta}^{j\prime}}\right\}. \tag{S.41}$$

We firstly focus on the first derivative on the right-hand side of (S.41). This term reduces to:

$$\frac{\partial \text{vec}\left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\tilde{X}_t^{j\prime}Y_t\right)}{\partial\widehat{\beta}^{j\prime}} = \frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\left\{\left(Y_t' \otimes I_n\right)\frac{\partial \text{vec}\left(\tilde{X}_t^{j\prime}\right)}{\partial\widehat{\beta}^{j\prime}}\right\},$$
$$\frac{\partial \text{vec}\left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\tilde{X}_t^{j\prime}Y_t\right)}{\partial\widehat{\beta}^{j\prime}} = \frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\left\{\left(Y_t' \otimes I_n\right)\left(I_K \otimes R'\right)\Big[\left(I_1 \otimes \mathbb{K}_{K,f} \otimes I_K\right) \times \right.$$
$$\left.\left(I_f \otimes \text{vec}\left(I_K\right)\right)\Big]\frac{\partial \text{vec}\left(\widehat{X}_t^j\right)}{\partial\widehat{\beta}^{j\prime}}\right\}, \tag{S.42}$$

where $\mathbb{K}_{K,f}$ accounts for the commutation matrix evaluated at K and f, and $f = K(p+q+1)$. The last term in (S.42) reduces to:

$$\frac{\partial \text{vec}\left(\widehat{X}_t^j\right)}{\partial\widehat{\beta}^{j\prime}} = \frac{\partial\left(\text{vec}\left(Y_t - \widehat{U}_t^j, Y_{t-1}, \cdots, Y_{t-p}, \widehat{U}_{t-1}^j, \cdots, \widehat{U}_{t-q}^j\right)\right)}{\partial\widehat{\beta}^{j\prime}},$$
$$\frac{\partial \text{vec}\left(\widehat{X}_t^j\right)}{\partial\widehat{\beta}^{j\prime}} = \text{vec}\left(-\frac{\partial\widehat{U}_t^j}{\partial\widehat{\beta}^{j\prime}}, 0_{K,n}, \cdots 0_{K,n}, \frac{\partial\widehat{U}_{t-1}^j}{\partial\widehat{\beta}^{j\prime}}, \cdots, \frac{\partial\widehat{U}_{t-q}^j}{\partial\widehat{\beta}^{j\prime}}\right). \tag{S.43}$$

Lemma 12.1 in Lütkepohl (2007) provides a closed recursive solution for the partial derivative in (S.43),

$$\frac{\partial \widehat{U}_t^j}{\partial \widehat{\beta}^{j\prime}} = \left\{ \left(\widehat{A}_0^{j^{-1}} \left[\widehat{A}_1^j Y_{t-1} + ... + \widehat{A}_p^j Y_{t-p} + \widehat{M}_1^j \widehat{U}_{t-1}^j + ... + \widehat{M}_q^j \widehat{U}_{t-q}^j \right] \right)' \otimes \widehat{A}_0^{j^{-1}} \right\} \times \left[I_{K^2} : 0 : ... : 0 \right] R - \left[\left(Y_{t-1}', ..., Y_{t-p}', \widehat{U}_{t-1}^{j\prime}, ..., \widehat{U}_{t-q}^{j\prime} \right) \otimes \widehat{A}_0^{j^{-1}} \right] \left[0 : I_{K^2(p+q)} \right] R - \widehat{A}_0^{j^{-1}} \left[M_1^j \frac{\partial \widehat{U}_{t-1}^j}{\partial \widehat{\beta}^{j\prime}} + ... + M_q^j \frac{\partial \widehat{U}_{t-q}^j}{\partial \widehat{\beta}^{j\prime}} \right]. \tag{S.44}$$

Hence, by plugging (S.42) into (S.41), and using the results derived in (S.43) and (S.44), (S.41) reduces to:

$$\widehat{V}_T\left(\widehat{\beta}^j\right) = \left\{ \left[I_1 \otimes \left(\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \tilde{X}_t^{j\prime} \tilde{X}_t^j \right)^{-1} \right] \frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \left\{ \left(Y_t' \otimes I_n \right) \left(I_K \otimes R' \right) \times \left[\left(I_1 \otimes \mathbb{K}_{K,f} \otimes I_K \right) \left(I_f \otimes \text{vec}\left(I_K \right) \right) \right] \frac{\partial \text{vec}\left(\widehat{X}_t^j \right)}{\partial \widehat{\beta}^{j\prime}} \right\} + \left[\left(\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \tilde{X}_t^{j\prime} Y_t \right)' \otimes I_n \right] \frac{\partial \text{vec}\left(\left(\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \tilde{X}_t^{j\prime} \tilde{X}_t^j \right)^{-1} \right)}{\partial \widehat{\beta}^{j\prime}} \right\}. \tag{S.45}$$

We now turn our attention to the last term on the right-hand side of (S.45). Define $Z^j = \left(\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \tilde{X}_t^{j\prime} \tilde{X}_t^j \right)$, such that the last derivative in (S.45) is given by:

$$\begin{aligned} \frac{\partial \text{vec}\left(\left(Z^j \right)^{-1} \right)}{\partial \widehat{\beta}^{j\prime}} &= \left[-\left(Z^j \right)^{-1} \otimes \left(Z^j \right)^{-1} \right] \frac{\partial \text{vec}\left(Z^j \right)}{\partial \widehat{\beta}^{j\prime}} \\ \frac{\partial \text{vec}\left(Z^j \right)}{\partial \widehat{\beta}^{j\prime}} &= \left[-\left(Z^j \right)^{-1} \otimes \left(Z^j \right)^{-1} \right] \left[\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \left\{ \left(I_{n^2} + \mathbb{K}_{n,n} \right) \times \left(I_n \otimes \tilde{X}_t^{j\prime} \right) \times \left(R' \otimes I_K \right) \left[\left(I_f \otimes \mathbb{K}_{K,1} \otimes I_K \right) \times \left(I_f \otimes \text{vec}\left(I_K \right) \right) \right] \frac{\partial \text{vec}\left(\widehat{X}_t^{j\prime} \right)}{\partial \widehat{\beta}^{j\prime}} \right\} \right]. \end{aligned} \tag{S.46}$$

Plugging (S.46) into (S.45) we obtain a closed solution for the sample gradient evaluated on any point satisfying Assumptions B.1,

$$\widehat{V}_T\left(\widehat{\beta}^j\right) = \left\{\left[I_1 \otimes Z^{j^{-1}}\right] \frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \left\{\left(Y_t' \otimes I_n\right)\left(I_K \otimes R'\right) \times \right.\right.$$
$$\left.\left[\left(I_1 \otimes \mathbb{K}_{K,f} \otimes I_K\right)\left(I_f \otimes \text{vec}\left(I_K\right)\right)\right] \frac{\partial \text{vec}\left(\widehat{X}_t^j\right)}{\partial \widehat{\beta}^{j\prime}}\right\} +$$
$$\left\{\left[\left(\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \tilde{X}_t^{j\prime} Y_t\right)' \otimes I_n\right]\left[-\left(Z^j\right)^{-1} \otimes \left(Z^j\right)^{-1}\right] \times \right. \tag{S.47}$$
$$\left[\frac{1}{T-\bar{q}} \sum_{t=1+\bar{q}}^{T} \left\{\left[\left(I_{n^2} + \mathbb{K}_{n,n}\right)\left(I_n \otimes \tilde{X}_t^{j\prime}\right)\left(R' \otimes I_K\right) \times \right.\right.\right.$$
$$\left.\left.\left.\left[\left(I_f \otimes \mathbb{K}_{K,1} \otimes I_K\right)\left(I_f \otimes \text{vec}\left(I_K\right)\right)\right] \frac{\partial \text{vec}\left(\widehat{X}_t^{j\prime}\right)}{\partial \widehat{\beta}^{j\prime}}\right\}\right]\right\}$$

where $\frac{\partial \text{vec}(\widehat{X}_t^j)}{\partial \widehat{\beta}^{j\prime}} = \frac{\partial \text{vec}(\widehat{X}_t^{j\prime})}{\partial \widehat{\beta}^{j\prime}}$ are obtained using (S.43) and (S.44).

S. 2.3 General Lemmas

Lemmas in this section nest both the univariate ARMA(1,1) and the general VARMA(p,q) models. We base the notation on the general VARMA(p,q) model, since it encompasses the univariate ARMA(1,1) model. Therefore, all Lemmas hold under appropriate lag order and system dimension. We refer to the set of Assumptions B as the necessary conditions that guarantee the Lemmas in this section hold. In particular, we assume Assumption B.2 rather than Assumption B.2a, however it is important to note that all the results in this section remain valid under Assumption B.2a, because Assumption B.2a implies that innovations are uncorrelated, allowing the use of the weak law of large numbers. Also, recall that the set of Assumptions B is the multivariate counterpart of the set of Assumptions A. This extends the validity of the Lemmas to the univariate ARMA(1,1) model. Throughout this section, $\bar{q} = \max\{p, q\}$ is the maximum lag order of the general VARMA(p,q) model; $\phi \in \mathbb{B}$ collects all the free parameters in the model such that Assumption B.1 is satisfied; $\|.\|$ accounts for the Frobenius norm; and $\breve{N}_T\left(\breve{\beta}^j\right)$ is the infeasible sample mapping at any j iteration defined as

$$\breve{N}_T\left(\breve{\beta}^j\right) = \left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}^{j\prime}_{\infty,t}\widetilde{X}^{j}_{\infty,t}\right]^{-1}\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}^{j\prime}_{\infty,t}Y_t\right], \tag{S.48}$$

where $\widetilde{X}^j_{\infty,t} = \left[\left(X^{j\prime}_{\infty,t}\otimes I_K\right)R\right]$ is a $(K\times n)$ matrix, with $X^j_{\infty,t} = \text{vec}\left(Y_t - U^j_t, Y_{t-1}, ..., Y_{t-p}, U^j_{t-1}, ..., U^j_{t-q}\right)$; and U^j_t is computed recursively using the VARMA(p,q) model in the fashion of (15), where all the initial values are assumed to be known, i.e., $Y_{t-\ell}$ and $U^j_{t-\ell}$ are known for all $\ell \geq t$. Similarly, $\widehat{N}_T\left(\widehat{\beta}^j\right)$ is the sample mapping at any j iteration defined as

$$\widehat{N}_T\left(\widehat{\beta}^j\right) = \left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}^{j\prime}_{t}\widetilde{X}^{j}_{t}\right]^{-1}\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}^{j\prime}_{t}Y_t\right], \tag{S.49}$$

where $\widetilde{X}^j_t = \left[\left(\widehat{X}^{j\prime}_t\otimes I_K\right)R\right]$ has dimension $(K\times n)$ and denote the regressors computed on the j^{th} iteration; $\widehat{X}^j_t = \text{vec}(Y_t - \widehat{U}^j_t, Y_{t-1}, ..., Y_{t-p}, \widehat{U}^j_{t-1}, ..., \widehat{U}^j_{t-q})$; and $\widehat{U}^j_t$ is computed recursively as in (15), where all the initial values are set to zero, i.e., $Y_{t-\ell} = \widehat{U}^j_{t-\ell} = 0$ for all $\ell \geq t$. Because Lemma 4 deals explicitly with the effect of initial values, without any loss of generality, we highlight that summations start from $\bar{q}+1$ in both infeasible and sample mappings.

Lemma 4 *Assume Assumptions B.1 and B.2 hold. Then,*
$\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi) - \breve{N}_T(\phi)\right\| = O_p\left(T^{-1}\right)$.

Proof of Lemma 4: Recall that Assumption B.1 (A.1 in the ARMA(1,1) case) guarantees that the set $\mathbb{B}$ is closed, such that the $A(L)$ and $M(L)$ polynomials implied by any element in the set $\mathbb{B}$ have absolute eigenvalues bounded by some constant ρ, satisfying $0<\rho<1$; Using the sample mapping for a general VARMA(p,q) as in (S.49), we bound the difference between sample mapping and its infeasible counterpart in (S.48) as

$$\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-\breve{N}_{\infty,T}(\phi)\right\| \leq \sup_{\phi\in\mathbb{B}}\left\{\left\|\left[Z^{j^{-1}}-Z_{\infty}^{j^{-1}}\right]\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}Y_t\right]\right\|+\right.$$
$$\left.\left\|Z_{\infty}^{j^{-1}}\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}Y_t-\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}Y_t\right]\right\|\right\},$$
$$\leq \sup_{\phi\in\mathbb{B}}\left\{\left\|\left[Z^{j^{-1}}-Z_{\infty}^{j^{-1}}\right]\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}Y_t\right]\right\|\right\}+ \tag{S.50}$$
$$\sup_{\phi\in\mathbb{B}}\left\{\left\|Z_{\infty}^{j^{-1}}\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}Y_t-\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}Y_t\right]\right\|\right\},$$

where $Z_{\infty}^{j}=\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}\widetilde{X}_{\infty,t}^{j}$ and $Z^{j}=\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}\widetilde{X}_t^{j}$. The first term on the right-hand side of (S.50) can be bounded as:

$$\sup_{\phi\in\mathbb{B}}\left\{\left\|\left[Z^{j^{-1}}-Z_{\infty}^{j^{-1}}\right]\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}Y_t\right]\right\|\right\}\leq \sup_{\phi\in\mathbb{B}}\left\{\left\|Z^{j^{-1}}-Z_{\infty}^{j^{-1}}\right\|\left\|\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}Y_t\right\|\right\}$$
$$\leq \sup_{\phi\in\mathbb{B}}\left\{\left\|Z^{j^{-1}}\right\|\left\|Z^{j}-Z_{\infty}^{j}\right\|\left\|-Z_{\infty}^{j^{-1}}\right\|\times\right.$$
$$\left.\left\|\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}Y_t\right\|\right\}. \tag{S.51}$$

The first, third and fourth terms on the right-hand side of (S.51) are $O_p(1)$ quantities, because they are the sample counterparts of second moment matrices obtained from covariance stationary VARMA(p,q), as $\phi\in\mathbb{B}$. To show that $\sup_{\phi\in\mathbb{B}}\left\|Z^{j}-Z_{\infty}^{j}\right\|=O_p\left(T^{-1}\right)$, we firstly bound this term as

$$\sup_{\phi\in\mathbb{B}}\left\|Z^{j}-Z_{\infty}^{j}\right\|=\sup_{\phi\in\mathbb{B}}\left\|\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_t^{j\prime}\widetilde{X}_t^{j}-\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}\widetilde{X}_{\infty,t}^{j}\right\|,$$
$$\leq \frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_t^{j\prime}\widetilde{X}_t^{j}-\widetilde{X}_{\infty,t}^{j\prime}\widetilde{X}_{\infty,t}^{j}\right\|,$$
$$\leq \frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\left\{\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_t^{j\prime}\right\|\left\|\widetilde{X}_t^{j}-\widetilde{X}_{\infty,t}^{j}\right\|+\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_t^{j\prime}-\widetilde{X}_{\infty,t}^{j\prime}\right\|\left\|\widetilde{X}_{\infty,t}^{j}\right\|\right\}. \tag{S.52}$$

$\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_{\infty,t}^{j}\right\|$ and $\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_t^{j\prime}\right\|$ are $O_p(1)$, as Assumption B.1 delivers stability and invertibility for the VARMA(p,q) model, such that

$$\sup_{\phi\in\mathbb{B}}\left\|Z^{j}-Z_{\infty}^{j}\right\|\leq\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\left\{O_{p}\left(1\right)\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_{t}^{j}-\widetilde{X}_{\infty,t}^{j}\right\|+\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_{t}^{j\prime}-\widetilde{X}_{\infty,t}^{j\prime}\right\|O_{p}\left(1\right)\right\}. \tag{S.53}$$

Bound $\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_{t}^{j}-\widetilde{X}_{\infty,t}^{j}\right\|$ as

$$\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_{t}^{j}-\widetilde{X}_{\infty,t}^{j}\right\|=\sup_{\phi\in\mathbb{B}}\left\|\left(\widehat{X}_{t}^{j\prime}\otimes I_{K}\right)R-\left(X_{\infty,t}^{j\prime}\otimes I_{K}\right)R\right\|, \tag{S.54}$$

$$\begin{aligned}
&\leq\sup_{\phi\in\mathbb{B}}\left\|\left(\widehat{X}_{t}^{j\prime}\otimes I_{K}\right)-\left(X_{\infty,t}^{j\prime}\otimes I_{K}\right)\right\|\left\|R\right\|,\\
&\leq\sup_{\phi\in\mathbb{B}}\left\|\left(\widehat{X}_{t}^{j\prime}-X_{\infty,t}^{j\prime}\right)\otimes I_{K}\right\|\left\|R\right\|=\sup_{\phi\in\mathbb{B}}\left\{\left\|\widehat{X}_{t}^{j\prime}-X_{\infty,t}^{j\prime}\right\|^{2}\left\|I_{K}\right\|^{2}\right\}^{\frac{1}{2}}\left\|R\right\|,\\
&\leq\sup_{\phi\in\mathbb{B}}\left\|\widehat{X}_{t}^{j\prime}-X_{\infty,t}^{j\prime}\right\|\left\|I_{K}\right\|\left\|R\right\|,\\
&\leq\sup_{\phi\in\mathbb{B}}\left\|\widehat{X}_{t}^{j\prime}-X_{\infty,t}^{j\prime}\right\|O\left(1\right)O\left(1\right),\\
&\leq\sup_{\phi\in\mathbb{B}}\left\|\widehat{X}_{t}^{j\prime}-X_{\infty,t}^{j\prime}\right\|O\left(1\right).
\end{aligned} \tag{S.55}$$

Recall that $\widehat{X}_{t}^{j}$ and $X_{\infty,t}^{j}$ are $(K(p+q+1)\times 1)$ vectors given by

$$\widehat{X}_{t}^{j}=\text{vec}\left(Y_{t}-\widehat{U}_{t}^{j},Y_{t-1},...,Y_{t-p},\widehat{U}_{t-1}^{j},...,\widehat{U}_{t-q}^{j}\right), \tag{S.56}$$

$$X_{\infty,t}^{j}=\text{vec}\left(Y_{t}-U_{t}^{j},Y_{t-1},...,Y_{t-p},U_{t-1}^{j},...,U_{t-q}^{j}\right). \tag{S.57}$$

Moreover, write a general VARMA(p,q) model as a VARMA(1,1) model as

$$\begin{pmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-\bar{q}+1} \end{pmatrix} = \begin{pmatrix} \widehat{A}_0^{j^{-1}} \widehat{A}_1^j & \widehat{A}_0^{j^{-1}} \widehat{A}_2^j & \dots & 0 & \widehat{A}_0^{j^{-1}} \widehat{A}_{\bar{q}}^j \\ I_K & 0 & \dots & 0 & 0 \\ 0 & I_K & \dots & 0 & 0 \\ 0 & 0 & \dots & I_K & 0 \end{pmatrix} \begin{pmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-\bar{q}} \end{pmatrix} + \begin{pmatrix} U_t^j \\ U_{t-1}^j \\ \vdots \\ U_{t-\bar{q}+1}^j \end{pmatrix} +$$

$$\begin{pmatrix} \widehat{A}_0^{j^{-1}} \widehat{M}_1^j & \widehat{A}_0^{j^{-1}} \widehat{M}_2^j & \dots & 0 & \widehat{A}_0^{j^{-1}} \widehat{M}_{\bar{q}}^j \\ -I_K & 0 & \dots & 0 & 0 \\ 0 & -I_K & \dots & 0 & 0 \\ 0 & 0 & \dots & -I_K & 0 \end{pmatrix} + \begin{pmatrix} U_{t-1}^j \\ U_{t-2}^j \\ \vdots \\ U_{t-\bar{q}}^j \end{pmatrix}, \tag{S.58}$$

with a compact notation given by

$$Y_{\bar{q},t} = \widehat{A}^j Y_{\bar{q},t-1} + U_{\bar{q},t}^j + \widehat{M}^j U_{\bar{q},t-1}^j, \tag{S.59}$$

where $\phi = \text{vec}\left(\widehat{A}_0^j, \widehat{A}_1^j, ..., \widehat{A}_{\bar{q}}^j, \widehat{M}_1^j, ..., \widehat{M}_{\bar{q}}^j\right)$; $Y_{\bar{q},t}$ and $U_{\bar{q},t}^j$ are $(K\bar{q} \times 1)$ vectors; and $\widehat{A}^j$ and $\widehat{M}^j$ are $(K\bar{q} \times K\bar{q})$ matrices. Using, (S.59), it follows that

$$\widehat{U}_t^j - U_t^j = R_{\bar{q}}\left[(-\widehat{M}^j)^{t-\bar{q}}\left(U_{\bar{q},\bar{q}}^j - \widehat{U}_{\bar{q},\bar{q}}^j\right)\right] = R_{\bar{q}}\left[(-\widehat{M}^j)^{t-\bar{q}} C_{\bar{q}}\right], \text{ for } t \geq \bar{q}, \tag{S.60}$$

$$\widehat{U}_t^j - U_t^j = \sum_{i=t}^{\bar{q}} \widehat{A}_i^j Y_{t-i} + \sum_{i=t}^{\bar{q}} \widehat{M}_i^j Y_{t-i} - \sum_{i=1}^{t-1} \widehat{M}_i^j \left(\widehat{U}_{t-i}^j - U_{t-i}^j\right) \mathbb{1}(t > 1), \text{ for } 1 \leq t < \bar{q}, \tag{S.61}$$

where $\mathbb{1}(t > 1)$ is the indicator function that returns 1 if $t > 1$; $R_{\bar{q}} = (I_K, 0, ..., 0)$ is a $(K \times K\bar{q})$ selection matrix; and $C_{\bar{q}} = U_{\bar{q},\bar{q}}^j - \widehat{U}_{\bar{q},\bar{q}}^j$. Note that because invertibility and stability conditions hold from Assumption B.1, $C_{\bar{q}}' C_{\bar{q}} < \infty$. Using (S.60),

$$\begin{aligned} \left\|\widehat{X}_t^j - X_{\infty,t}^j\right\| &= \text{tr}\left[\left(\widehat{X}_t^j - X_{\infty,t}^j\right)\left(\widehat{X}_t^{j\prime} - X_{\infty,t}^{j\prime}\right)'\right]^{1/2} \\ &= \text{tr}\left[\sum_{i=0}^{\bar{q}} \left(R_{\bar{q}}(-\widehat{M}^j)^{(t-\bar{q}-i)} C_{\bar{q}}\right)' \left(R_{\bar{q}}(-\widehat{M}^j)^{(t-\bar{q}-i)} C_{\bar{q}}\right)\right]^{1/2}, \text{ for } t \geq 2\bar{q}. \end{aligned} \tag{S.62}$$

Assumption B.1 imposes that both the general VARMA(p,q) and its VARMA(1,1) representation are invertible, implying that the maximum eigenvalue of $\widehat{M}^j$ in absolute value is bounded away from 1 for any $\phi \in \mathbb{B}$ and $j \geq 0$, such that $\left\|\widehat{M}^{t-\bar{q}}\right\| \leq \bar{C}\rho^{t-\bar{q}}$, with $\rho \in (0,1)$ and $\bar{C} > 0$ being a finite real constant. Applying this result to (S.62), we have that

$$\sup_{\phi \in \mathbb{B}} \left\|\widetilde{X}_t^j - \widetilde{X}_{\infty,t}^j\right\| = O_p\left(\rho^{t-2\bar{q}}\right), \quad \text{for } t \geq 2\bar{q}. \tag{S.63}$$

Using (S.60), we have that for $\bar{q}+1 \leq t < 2\bar{q}$,

$$\left\|\widetilde{X}_t^j - \widetilde{X}_{\infty,t}^j\right\| = \left[\sum_{i=t-\bar{q}}^{\bar{q}} \left(\widehat{U}_t^j - U_t^j\right)' \left(\widehat{U}_t^j - U_t^j\right) + \sum_{i=\bar{q}+1}^{t} \left(R_{\bar{q}}\left(-\widehat{M}^j\right)^{i-\bar{q}} C_{\bar{q}}\right)' \left(R_{\bar{q}}\left(-\widehat{M}^j\right)^{i-\bar{q}} C_{\bar{q}}\right)\right]^{1/2}, \tag{S.64}$$

which implies that $\sup_{\phi \in \mathbb{B}} \left\|\widetilde{X}_t^j - \widetilde{X}_{\infty,t}^j\right\| = O_p(1)$ for $\bar{q}+1 \leq t < 2\bar{q}$, as (S.64) is a finite sum of bounded terms. Thus, it follows from (S.63) and (S.64), that there exist a constant C such that

$$\sup_{\phi \in \mathbb{B}} \left\|Z^j - Z_\infty^j\right\| \leq \frac{1}{T-\bar{q}}\left\{O_p(1) + C\sum_{t=2\bar{q}}^{T} \rho^{t-2\bar{q}}\right\}.$$

As $T \longrightarrow \infty$, we finally have that

$$\sup_{\phi \in \mathbb{B}} \left\|Z^j - Z_\infty^j\right\| = O_p\left(T^{-1}\right), \tag{S.65}$$

implying that the first term of (S.50) has order $O_p\left(T^{-1}\right)$.

We follow similar steps to show that the second term of (S.50) is $O_p\left(T^{-1}\right)$. In particular,

$$\sup_{\phi \in \mathbb{B}} \left\{\left\|Z_\infty^{j^{-1}}\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T} \widetilde{X}_t^{j\prime} Y_t - \frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T} \widetilde{X}_{\infty,t}^{j\prime} Y_t\right]\right\|\right\} \leq \sup_{\phi \in \mathbb{B}} \left\|Z_\infty^{j^{-1}}\right\| \times \frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\left\{\sup_{\phi \in \mathbb{B}}\left\|\widetilde{X}_t^{j\prime} - \widetilde{X}_{\infty,t}^{j\prime}\right\| \sup_{\phi \in \mathbb{B}} \|Y_t\|\right\}. \tag{S.66}$$

Because $\phi \in \mathbb{B}$ satisfies Assumption B.1, we have that $\sup_{\phi \in \mathbb{B}} \|Y_t\| = O_p(1)$ and $\sup_{\phi \in \mathbb{B}} \left\|Z_\infty^{j^{-1}}\right\| =$

$O_p(1)$. Using (S.63) and (S.64), we have that

$$\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\left\{\sup_{\phi\in\mathbb{B}}\left\|\widetilde{X}_t^{j\prime}-\widetilde{X}_{\infty,t}^{j\prime}\right\|\sup_{\phi\in\mathbb{B}}\|Y_t\|\right\}\leq\frac{1}{T-\bar{q}}\left(O_p(1)+C\sum_{t=2\bar{q}}^{T}\rho^{t-2\bar{q}}\right). \tag{S.67}$$

Hence, using similar arguments as in (S.65), we find that (S.67) is $O_p\left(T^{-1}\right)$ as $T\longrightarrow\infty$, implying that the second term of (S.50) is $O_p\left(T^{-1}\right)$. Therefore,

$$\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-\breve{N}_T(\phi)\right\|=O_p\left(T^{-1}\right). \tag{S.68}$$

Lemma 5 *Assume Assumptions B.1, B.2 and B.3 hold. Then,* $\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-N(\phi)\right\|=o_p(1)$ *as* $T\longrightarrow\infty$.

Proof of Lemma 5: First, note that

$$\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-N(\phi)\right\|\leq\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-\breve{N}_T(\phi)\right\|+\sup_{\phi\in\mathbb{B}}\left\|\breve{N}_T(\phi)-N(\phi)\right\|. \tag{S.69}$$

Lemma 4 gives that the first term on the right-hand side of (S.69) has order $o_p(1)$ as $T\longrightarrow\infty$. Bound the second term on the right-hand side of (S.69) as

$$\begin{aligned}\sup_{\phi\in\mathbb{B}}\left\|\breve{N}_T(\phi)-N(\phi)\right\|\leq&\sup_{\phi\in\mathbb{B}}\left\{\left\|Z_\infty^{j^{-1}}\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}Y_t-\mathbb{E}\left[\widetilde{X}_{\infty,t}^{j\prime}Y_t\right]\right]\right\|+\right.\\&\left.\left\|\left[Z_\infty^{j^{-1}}-\mathbb{E}\left[Z_{\infty,t}^{j}\right]^{-1}\right]\mathbb{E}\left[\widetilde{X}_{\infty,t}^{j\prime}Y_t\right]\right\|\right\},\\\leq&\sup_{\phi\in\mathbb{B}}\left\{\left\|Z_\infty^{j^{-1}}\right\|\left\|\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}Y_t-\mathbb{E}\left[\widetilde{X}_{\infty,t}^{j\prime}Y_t\right]\right]\right\|\right\}+\\&\sup_{\phi\in\mathbb{B}}\left\{\left\|Z_\infty^{j^{-1}}\right\|\left\|Z_\infty^{j}-\mathbb{E}\left[Z_{\infty,t}^{j}\right]\right\|\left\|\mathbb{E}\left[Z_{\infty,t}^{j}\right]^{-1}\right\|\left\|\mathbb{E}\left[\widetilde{X}_{\infty,t}^{j\prime}Y_t\right]\right\|\right\},\end{aligned} \tag{S.70}$$

where $Z_\infty^j=\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}_{\infty,t}^{j\prime}\widetilde{X}_{\infty,t}^{j}$ and $Z_{\infty,t}^{j}=\widetilde{X}_{\infty,t}^{j\prime}\widetilde{X}_{\infty,t}^{j}$. The first, third, fifth, and sixth terms on the right-hand side of (S.70) are $O_p(1)$, because $\phi\in\mathbb{B}$ implies that Assumption

B.1 is satisfied,

$$\sup_{\phi\in\mathbb{B}}\left\|\breve{N}_T(\phi)-N(\phi)\right\|\leq O_p(1)\sup_{\phi\in\mathbb{B}}\left\|\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}^{j\prime}_{\infty,t}Y_t-\mathbb{E}\left[\widetilde{X}^{j\prime}_{\infty,t}Y_t\right]\right]\right\|+ \\ O_p(1)\sup_{\phi\in\mathbb{B}}\left\{\left\|Z^j_{\infty}-\mathbb{E}\left[Z^j_{\infty,t}\right]\right\|\right\}O_p(1)\,O_p(1)\,. \tag{S.71}$$

We prove that $\sup_{\phi\in\mathbb{B}}\left\|\breve{N}_T(\phi)-N(\phi)\right\|$ has order $o_p(1)$ in two steps. We first prove that $\left\|\breve{N}_T(\phi)-N(\phi)\right\|$ is point-wise $o_p(1)$. To this end, recall that $\widetilde{X}^{j\prime}_{\infty,t}=R'\left(X^j_{\infty,t}\otimes I_K\right)$, and R is a deterministic matrix containing ones and zeros. Point-wise convergence of $\left\|\breve{N}_T(\phi)-N(\phi)\right\|$ becomes then a law of large numbers problem, where it suffices to show that:

$$\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\left(X^j_{\infty,t}\otimes I_K\right)Y_t\xrightarrow{p}\mathbb{E}\left[\left(X^j_{\infty,t}\otimes I_K\right)Y_t\right], \tag{S.72}$$

$$\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}X^j_{\infty,t}X^{j\prime}_{\infty,t}\xrightarrow{p}\mathbb{E}\left[X^j_{\infty,t}X^{j\prime}_{\infty,t}\right]. \tag{S.73}$$

Assumptions B.1 and B.2a guarantee that the VARMA(p,q) model is covariance-stationary[1]. This allows us to use the weak law of large numbers, such that (S.72) and (S.73) hold for each $\phi\in\mathbb{B}$ as $T\longrightarrow\infty$, implying that

$$\left\|\breve{N}_T(\phi)-N(\phi)\right\|=o_p(1) \tag{S.74}$$

for each $\phi\in\mathbb{B}$. The second step consists on establishing the uniform convergence, i.e., that $\sup_{\phi\in\mathbb{B}}\left\|\breve{N}_T(\phi)-N(\phi)\right\|$ has order $o_p(1)$. Using Theorem 21.9 Davidson, 1994, pg. 337, uniform convergence arises if $\left\|\breve{N}_T(\phi)-N(\phi)\right\|$ converges point-wise to zero for each $\phi\in\mathbb{B}$, and $\breve{N}_T(\phi)$ is stochastically equicontinuous for all $\phi\in\mathbb{B}$. For the univariate ARMA(1,1) specification, Lemma 3 guarantees that $\breve{N}_T(\phi)$ is stochastically equicontinuous. For the general VARMA(p,q) model, the existence of Lipschitz-type of condition as in item ii. in Assumption B.3 suffices to deliver stochastic equicontinuity (see Lemma 2.9 in Newey and McFadden, 1994, pg. 2138). Hence, using Theorem 21.9 Davidson, 1994, pg. 337, such that we combine point-wise convergence given in (S.74) with $\breve{N}_T(\phi)$ being stochastically equicontinuous for all $\phi\in\mathbb{B}$ suffice to yield that $\sup_{\phi\in\mathbb{B}}\left\|\breve{N}_T(\phi)-N(\phi)\right\|$ has order $o_p(1)$,

[1]For the univariate ARMA(1,1), Assumptions A.1 and B.2 deliver covariance stationarity.

and therefore proves this Lemma.

Lemma 6 *Assume Assumptions B.1, B.2 and B.3 hold. Then,* $\sup_{\phi,\gamma\in\mathbb{B}}\left\|\left[\widehat{\Lambda}_T(\phi,\gamma)-\Lambda(\phi,\gamma)\right](\phi-\gamma)\right\| = o_p(1)$ *as* $T\longrightarrow\infty$.

Proof of Lemma 6: This proof follows the steps in Dominitz and Sherman (2005). Fix $\phi,\gamma\in\mathbb{B}$, such that Assumption B.1 is satisfied. Using the mean value theorem, rewrite the difference between the population and sample mappings, evaluated at different vector of estimates, as

$$N(\phi)-N(\gamma)=\Lambda(\phi,\gamma)[\phi-\gamma], \tag{S.75}$$

$$\widehat{N}_T(\phi)-\widehat{N}_T(\gamma)=\widehat{\Lambda}_T(\phi,\gamma)[\phi-\gamma], \tag{S.76}$$

with $\Lambda(\phi,\gamma)=\int_0^1 V(\phi+\xi(\phi-\gamma))\,d\xi$ and $\widehat{\Lambda}_T(\phi,\gamma)=\int_0^1 \widehat{V}_T(\phi+\xi(\phi-\gamma))\,d\xi$. Using (S.76) and (S.75), rewrite $\sup_{\phi,\gamma\in\mathbb{B}}\left\|\left[\widehat{\Lambda}_T(\phi,\gamma)-\Lambda(\phi,\gamma)\right](\phi-\gamma)\right\|$ as

$$\sup_{\phi,\gamma\in\mathbb{B}}\left\|\left[\widehat{\Lambda}_T(\phi,\gamma)-\Lambda(\phi,\gamma)\right](\phi-\gamma)\right\| = \sup_{\phi,\gamma\in\mathbb{B}}\left\|\left[\widehat{N}_T(\phi)-\widehat{N}_T(\gamma)\right]-\left[N_T(\phi)-N_T(\gamma)\right]\right\|,$$

$$\begin{aligned}\sup_{\phi,\gamma\in\mathbb{B}}\left\|\left[\widehat{\Lambda}_T(\phi,\gamma)-\Lambda(\phi,\gamma)\right](\phi-\gamma)\right\| \leq{}& \sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-\breve{N}_T(\phi)\right\|+\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-\breve{N}_T(\phi)\right\|+\\ & \sup_{\gamma\in\mathbb{B}}\left\|\breve{N}_T(\gamma)-N(\gamma)\right\|+\sup_{\gamma\in\mathbb{B}}\left\|\breve{N}_T(\gamma)-N(\gamma)\right\|.\end{aligned} \tag{S.77}$$

Lemma 4 gives that the first two suprema on the right-hand side of (S.77) have order $O_p\left(T^{-1}\right)$, whereas it follows from Lemma 5 that the last two terms on the right-hand side of (S.77) have order $o_p(1)$ as $T\longrightarrow\infty$. These imply that

$$\sup_{\phi,\gamma\in\mathbb{B}}\left\|\left[\widehat{\Lambda}_T(\phi,\gamma)-\Lambda(\phi,\gamma)\right](\phi-\gamma)\right\| = o_p(1), \tag{S.78}$$

as $T\longrightarrow\infty$.

Lemma 7 *Assume Assumptions B.1, B.2 and B.3 hold. If*

i. $\sup_{\phi\in\mathbb{B}}\left|\widehat{N}_T(\phi)-N(\phi)\right| = o_p(1)$ *as* $T\longrightarrow\infty$

ii. $\sup_{\phi,\gamma\in\mathbb{B}}\left|\widehat{\Lambda}_T(\phi,\gamma)-\Lambda(\phi,\gamma)\right| = o_p(1)$ *as* $T\longrightarrow\infty$

then, $\widehat{N}_T(\phi)$ *is an ACM on* $(\mathbb{B}, E_n)$*, with* $\phi \in \mathbb{B}$ *and it has fixed point denoted by* $\widehat{\beta}$*, such that* $\left|\widehat{\beta}^j - \widehat{\beta}\right| = o_p(1)$ *uniformly as* $j, T \longrightarrow \infty$*.*

Proof of Lemma 7: Provided that $N(\phi)$ is an ACM on $(\mathbb{B}, E_n)$, with $\phi \in \mathbb{B}$, we have that $|N(\phi) - N(\gamma)| \leq \kappa |\phi - \gamma|$ holds for each $\phi, \gamma \in \mathbb{B}$. Following that, we bound $\left|\widehat{N}_T(\phi) - \widehat{N}_T(\gamma)\right|$ as:

$$\left|\widehat{N}_T(\phi) - \widehat{N}_T(\gamma)\right| \leq |N(\phi) - N(\gamma)| + \left|\left[\widehat{N}_T(\phi) - \widehat{N}_T(\gamma)\right] - [N(\phi) - N(\gamma)]\right|, \tag{S.79}$$

$$\left|\widehat{N}_T(\phi) - \widehat{N}_T(\gamma)\right| \leq \kappa |\phi - \gamma| + \left|\left[\widehat{\Lambda}_T(\phi, \gamma) - \Lambda(\phi, \gamma)\right][\phi - \gamma]\right|. \tag{S.80}$$

From Lemma 6, the second term on the right-hand of equation (S.80) has order $o_p(1)$. Thus as $T \longrightarrow \infty$, we have that $\left|\widehat{N}_T(\phi) - \widehat{N}_T(\gamma)\right| \leq \kappa |\phi - \gamma|$ yielding the first result of Lemma 7. The second step of the proof consists of showing that $\widehat{\beta}^j$ converges uniformly to the fixed point $\widehat{\beta}$ as $j, T \longrightarrow \infty$. To this purpose, because $\widehat{N}_T(\phi)$ with $\phi \in \mathbb{B}$ is an ACM with a fixed point $\widehat{\beta}$, such that $\widehat{\beta} = \widehat{N}_T\left(\widehat{\beta}\right)$, rewrite $\left|\widehat{\beta}^j - \widehat{\beta}\right|$ as

$$\begin{aligned}
\left|\widehat{\beta}^j - \widehat{\beta}\right| &= \left|\widehat{N}_T\left(\widehat{\beta}^{j-1}\right) - \widehat{N}_T\left(\widehat{\beta}\right)\right|, \\
&\leq \left|\widehat{N}_T\left(\widehat{\beta}^{j-1}\right) - N\left(\widehat{\beta}^{j-1}\right)\right| + \left|N\left(\widehat{\beta}^{j-1}\right) - \widehat{N}_T\left(\widehat{\beta}\right)\right|, \\
&\leq \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| + \left|N\left(\widehat{\beta}^{j-1}\right) - N(\beta)\right| + \left|N(\beta) - \widehat{N}_T\left(\widehat{\beta}\right)\right|, \\
&\leq \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| + \kappa \left|\widehat{\beta}^{j-1} - \beta\right| + \left|N(\beta) - \widehat{N}_T\left(\widehat{\beta}\right)\right|, \\
&\leq \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| + \kappa \left|\widehat{\beta}^{j-1} - \beta\right| + \left\{\left|N(\beta) - N\left(\beta^{j-1}\right)\right| + \left|N\left(\beta^{j-1}\right) - \widehat{N}_T\left(\widehat{\beta}\right)\right|\right\} \\
&\leq \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| + \kappa^j \left|\widehat{\beta}^0 - \beta\right| + \kappa^j \left|\beta - \beta^0\right| + \left|\beta^j - \widehat{\beta}\right| \\
&\leq \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| + \kappa^j \left\{\left|\widehat{\beta}^0 - \beta\right| + \left|\beta - \beta^0\right|\right\} + \left|\beta^j - \widehat{\beta}\right|. && \text{(S.81)}
\end{aligned}$$

The last term in (S.81) can be bounded as

$$\begin{aligned}
\left|\beta^j - \widehat{\beta}\right| &\leq \left|N\left(\beta^{j-1}\right) - N\left(\widehat{\beta}\right)\right| + \left|N\left(\widehat{\beta}\right) - \widehat{N}_T\left(\widehat{\beta}\right)\right| \\
&\leq \kappa \left|\beta^{j-1} - \widehat{\beta}\right| + \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right|. && \text{(S.82)}
\end{aligned}$$

Applying the inequality (S.82) recursively, $\left|\beta^j - \widehat{\beta}\right|$ reduces to

$$\left|\beta^j - \widehat{\beta}\right| \leq \kappa^j \left|\beta^0 - \widehat{\beta}\right| + \left(\sum_{i=0}^{j-1}\right) \kappa^i \left[\sup_{\phi\in\mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right|\right]. \tag{S.83}$$

Plug (S.83) into (S.81) and rearrange terms, so that

$$\left|\widehat{\beta}^j - \widehat{\beta}\right| \leq \sup_{\phi\in\mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| + \kappa^j \left\{\left|\widehat{\beta}^0 - \beta\right| + \left|\beta - \beta^0\right| + \left|\beta^0 - \widehat{\beta}\right|\right\} + \left(\sum_{i=0}^{j-1}\right) \kappa^i \left[\sup_{\phi\in\mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right|\right]. \tag{S.84}$$

As $j \longrightarrow \infty$ and provided that $\kappa \in (0, 1]$, the second term on the right-hand side of (S.84) is $o(1)$. This follows because $\left\{\widehat{\beta}^0, \beta^0, \beta, \widehat{\beta}\right\} \in \mathbb{B}$ and $\mathbb{B}$ is a closed ball centered at β, so that $\left\{\left|\widehat{\beta}^0 - \beta\right| + \left|\beta - \beta^0\right| + \left|\beta^0 - \widehat{\beta}\right|\right\}$ is $O(1)$. It follows that (S.84) reduces to

$$\left|\widehat{\beta}^j - \widehat{\beta}\right| \leq \sup_{\phi\in\mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right| + \left[\frac{1}{1-\kappa}\right] \sup_{\phi\in\mathbb{B}} \left|\widehat{N}_T(\phi) - N(\phi)\right|. \tag{S.85}$$

From Lemma 5 and because $\frac{1}{1-\kappa}$ is $O(1)$, the first and second terms on the right-hand side of (S.85) are uniformly $o_p(1)$. Therefore, $\left|\widehat{\beta}^j - \widehat{\beta}\right|$ is uniformly $o_p(1)$ as $j, T \longrightarrow \infty$.

Lemma 8 *Assume Assumptions B.1, B.2 and B.3 hold. If*

i. $\widehat{N}_T(\phi)$ *is an ACM on* $(\mathbb{B}, E_n)$

then, $\sqrt{T}\left|\widehat{\beta}^j - \widehat{\beta}\right| = o_p(1)$ *as* $j \longrightarrow \infty$ *with* $T \longrightarrow \infty$.

Proof of Lemma 8: We show the $\sqrt{T}$ convergence of $\widehat{\beta}^j$ to the fixed point $\widehat{\beta}$ by using the result that yields that the sample mapping is an ACM on $(\mathbb{B}, E_n)$. Denote $\widehat{\kappa}$ as the sample counterpart of κ. Then,

$$\sqrt{T}\left|\widehat{\beta}^j - \widehat{\beta}\right| = \sqrt{T}\left|\widehat{N}_T\left(\widehat{\beta}^{j-1}\right) - \widehat{N}_T\left(\widehat{\beta}\right)\right| \leq \sqrt{T}\left[\hat{\kappa}\left|\widehat{\beta}^{j-1} - \widehat{\beta}\right|\right]. \tag{S.86}$$

Substituting recursively (S.87), we have

$$\sqrt{T}\left|\widehat{\beta}^j - \widehat{\beta}\right| \leq \sqrt{T}\left[\widehat{\kappa}^j \left|\widehat{\beta}^0 - \widehat{\beta}\right|\right]. \tag{S.87}$$

To make the right-hand side of (S.87) converge in probability to zero, we require that $\hat{\kappa}^j$ dominates $\sqrt{T}$ as $j \longrightarrow \infty$ with $T \longrightarrow \infty$. A sufficient rate implying this dominance is one such that $j \gg -\frac{1}{2}\left[\frac{\ln(T)}{\ln(\kappa)}\right]$. Hence, provided that $\frac{\ln(T)}{j} = o\,(1)$, we have that $\sqrt{T}\left|\widehat{\beta}^{j} - \widehat{\beta}\right| = o_p\,(1)$, which proves the Lemma.

S. 3 Identification

This section discusses the Echelon form transformation and provides examples covering DGPs III, IV and V. We focus on these DGPs because they are the ones we extensively adopt in both Monte Carlo and Empirical studies. Recall from Section 2.1 that a stable and invertible VARMA(p,q) model is said to be in Echelon form if the conditions stated in equations (4), (5), (6) and (7) in the paper are satisfied (see Lütkepohl, 2007, pg. 452 and Lütkepohl and Poskitt (1996) for more details).

Example 1. (DGP III) Consider a stable and invertible nonstandard VARMA process with $K = 3$ and $\mathbf{p} = (1, 0, 0)'$. The resulting VARMA representation expressed in Echelon form is given by

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ a_{21,0} & 1 & 0 \\ a_{31,0} & 0 & 1 \end{pmatrix}}_{=A_0} Y_t = \underbrace{\begin{pmatrix} a_{11,1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{=A_1} Y_{t-1} + \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ a_{21,0} & 1 & 0 \\ a_{31,0} & 0 & 1 \end{pmatrix}}_{=A_0} U_t + \underbrace{\begin{pmatrix} m_{11,1} & m_{12,1} & m_{13,1} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{=M_1} U_{t-1}. \tag{S.88}$$

As noted in Hannan and Deistler (1988) and Lütkepohl and Poskitt (1996), the restrictions imposed by the Echelon form are necessary and sufficient for the unique identification of stable and invertible nonstandard VARMA models. This follows because the Kronecker index in the k^{th} row of $[A(L) : M(L)]$ only specifies the maximum degree of all operators, and hence further restrictions (potentially data dependent) could be added, (e.g. $a_{21,0} = a_{31,0} = m_{12,1} = m_{13,1} = 0$). The necessary and sufficient conditions implied by the $\mathbf{p}_{100}$ specification in (S.88) account for the most restrictive (lower number of free parameters) unique VARMA model we consider in this article. Multiplying both sides of (S.88) by A_0^{-1}, the nonstandard VARMA representation reduces to

$$Y_t = \begin{pmatrix} a^*_{11,1} & 0 & 0 \\ a^*_{21,1} & 0 & 0 \\ a^*_{31,1} & 0 & 0 \end{pmatrix} Y_{t-1} + U_t + \begin{pmatrix} m^*_{11,1} & m^*_{12,1} & m^*_{13,1} \\ m^*_{21,1} & m^*_{22,1} & m^*_{23,1} \\ m^*_{31,1} & m^*_{32,1} & m^*_{33,1} \end{pmatrix} U_{t-1}, \tag{S.89}$$

where $A_1^* = A_0^{-1}A_1$ and $M_1^* = A_0^{-1}M_1$. It is also clear from (S.89) that the necessary and sufficient restrictions imposed by the Echelon form do not allow DGP III to be partitioned into smaller independent systems. Recall from Section 2.2 that, when estimating the free parameters in (S.88), we augment the matrix of regressors with $Y_t - U_t = A_1^* Y_{t-1} + M_1^* U_{t-1}$. Therefore, because A_0 is different from an identity matrix and hence M_1^* is a full matrix, we have that $Y_t - U_t$ is a function of the K lagged innovations in (S.89). Invertibility of the MA polynomial, $M^*(L) = (I_K + M_1^* L)$, in (S.89) yields a VAR(∞) representation

$$Y_t = \sum_{i=1}^{\infty} \Pi_i Y_{t-i} + U_t, \tag{S.90}$$

where $\Pi_i = (-1)^{i-1}\left(M_1^{*i} + M_1^{*i-1}A_1^*\right)$, for $i = 1, 2, \ldots$. The AR parameters are now full parameter matrices with rank one, which allows us to rewrite (S.90) as

$$Y_t = \alpha \left(\sum_{i=1}^{\infty} b_i' Y_{t-i} \right) + U_t, \tag{S.91}$$

where α and b_i are $(K \times 1)$ vectors. Note that (S.91) is the infinite version of the RRVAR(p) model discussed in Carriero et al. (2011). Therefore, by specifying the Kronecker indices as $\mathbf{p} = (1, 0, 0)'$, we are able to generate a representation that is rich enough to be a function of the K innovations in the system. Furthermore, it yields a RRVAR(∞) model with coefficients that are functions of the free parameters in (S.88).

Using the restrictions imposed by the Echelon form transformation, it is possible to generalize the findings in Example 1 to the case where the first k Kronecker indices are equal to one and the remaining $K - k$ indices are zero. Example 2 nests DGP III, IV and V for $k = 1$, $k = 2$ and $k = 3$, respectively.

Example 2. Consider a K dimensional stable and invertible nonstandard VARMA process with Kronecker indices given by $\mathbf{p} = (p_1, p_2, \ldots, p_k, p_{k+1}, \ldots, p_K)'$, such that $p_i = 1$ for $i \leq k$

and $p_i = 0$ for $i > k$. The resulting VARMA representation expressed in Echelon form is given by

$$\underbrace{\begin{pmatrix} I_k & 0 \\ A_{0,K-k\times k} & I_{K-k} \end{pmatrix}}_{=A_0} Y_t = \underbrace{\begin{pmatrix} A_{1,k\times k} & 0 \\ 0 & 0 \end{pmatrix}}_{=A_1} Y_{t-1} + \underbrace{\begin{pmatrix} I_k & 0 \\ A_{0,K-k\times k} & I_{K-k} \end{pmatrix}}_{=A_0} U_t + \underbrace{\begin{pmatrix} M_{1,k\times K} \\ 0 \end{pmatrix}}_{=M_1} U_{t-1}. \tag{S.92}$$

Using similar arguments as in the Example 1, the standard representation of (S.92) reduces to

$$Y_t = A_1^* Y_{t-1} + U_t + M_1^* U_{t-1}, \tag{S.93}$$

where $A_1^* = \left(A_{1,K\times k}^*, 0\right)$ and M_1^* is a full matrix. The standard representation in (S.93) allows a RRVAR(∞) representation with parameter matrices of rank k. It follows, therefore, that

$$Y_t = \alpha \left(\sum_{i=1}^{\infty} b_i' Y_{t-i} \right) + U_t, \tag{S.94}$$

where α and b_i are $(K \times k)$ matrices.

From a practitioner's point of view, imposing uniqueness to stable and invertible non-standard VARMA(1,1) models using the Kronecker indices structure as in Example 2 is equivalent to allow the DGP to be an RRVAR(∞) model with rank k. A different interpretation follows from observing that (S.93) imposes that the K variables in the system depend on the lagged values of the first k variables (AR dynamics). With regard to the MA dynamics, the K variables in the system are allowed to respond to the K innovations in the system, following that M_1^* is a full matrix. Therefore, a practitioner could order the variables in such a way the first k variables would drive the AR dynamics in the system, while still allowing all shocks to play a role on the MA component. This type of strategy resembles the one used in the reduced form VAR framework to orthogonalize the impulse response functions by means of the Cholesky decomposition.

S. 4 Efficiency Loss

This section briefly discusses the asymptotic efficiency loss attached to the IOLS estimator compared to the efficient MLE estimator. We provide results considering the univariate ARMA(1,1) model. Using item (i.) in Theorem 1, the asymptotic variance of the IOLS estimator is given by

$$\Sigma_{\beta,IOLS} = \begin{pmatrix} -\frac{\left(-1+\beta_1^2\right)\left(1+2\beta_1\beta_2+\beta_2^2\right)}{(\beta_1+\beta_2)^2} & \frac{\left(-1+\beta_1^2\right)(1+\beta_1\beta_2)}{(\beta_1+\beta_2)^2} \\ \frac{\left(-1+\beta_1^2\right)(1+\beta_1\beta_2)}{(\beta_1+\beta_2)^2} & \frac{(1+\beta_1\beta_2)^2}{(\beta_1+\beta_2)^2} \end{pmatrix}, \tag{S.95}$$

while, as in Brockwell and Davis, 1987, pg. 253, the efficient asymptotic variance of the MLE estimator assumes the form of

$$\Sigma_{\beta,MLE} = \begin{pmatrix} \frac{-(-1+\beta_1^2)(1+\beta_1\beta_2)^2}{(\beta_1+\beta_2)^2} & -\frac{(-1+\beta_1^2)(1+\beta_1\beta_2)(-1+\beta_2^2)}{(\beta_1+\beta_2)^2} \\ -\frac{(-1+\beta_1^2)(1+\beta_1\beta_2)(-1+\beta_2^2)}{(\beta_1+\beta_2)^2} & -\frac{(1+\beta_1\beta_2)^2(-1+\beta_2^2)}{(\beta_1+\beta_2)^2} \end{pmatrix}. \tag{S.96}$$

Using (S.96) and (S.95), we obtain a theoretical expression for the difference between the asymptotic variance of the IOLS and MLE estimators,

$$W := \Sigma_{\beta,IOLS} - \Sigma_{\beta,MLE} = \begin{pmatrix} \frac{(-1+\beta_1^2)^2\beta_2^2}{(\beta_1+\beta_2)^2} & \frac{(-1+\beta_1^2)\beta_2^2(1+\beta_1\beta_2)}{(\beta_1+\beta_2)^2} \\ \frac{(-1+\beta_1^2)\beta_2^2(1+\beta_1\beta_2)}{(\beta_1+\beta_2)^2} & \frac{\beta_2^2(1+\beta_1\beta_2)^2}{(\beta_1+\beta_2)^2} \end{pmatrix}. \tag{S.97}$$

Note that the eigenvalues of W are given by $\left(0, \frac{\beta_2^2(2+\beta_1^4+2\beta_1\beta_2+\beta_1^2(-2+\beta_2^2))}{(\beta_1+\beta_2)^2}\right)'$, and positive semidefiniteness of W follows from $1 + \frac{\beta_1^4}{2} + \frac{\beta_1^2\beta_2^2}{2} + \beta_1^2 > \beta_1\beta_2$ for all parameters satisfying Assumption A.1. The positive semidefinite matrix W gives the efficiency loss of the IOLS estimator with respect to the MLE estimator. The behaviour of matrix W is closely related to the signs associated with β_1 and β_2. In one hand, when β_1 and β_2 have opposite signs and $(\beta_1 + \beta_2)$ approaches zero, all elements of W grow exponentially fast, following the fact that the denominator $(\beta_1 + \beta_2)^2$ is common to all elements of W. Using the model specification in (17), $(\beta_1 + \beta_2)$ approaches zero when the model is close to have common factors. In the other hand, when both parameters share the same sign, the diagonal elements of W are positive and bounded above by one. Finally, as an intuitive exercise, we can easily see that in the extreme case where $\beta_2 = 0$, the benchmark model becomes an AR(1) process where the matrix W is a zero matrix and the IOLS estimator converges at the first iteration ($V(\beta) = 0$).

S. 5 Monte Carlo Design

This section brings additional Monte Carlo results and therefore complements the selected tables reported in the paper. These additional simulations shed light on the relative finite sample performance of the IOLS estimator under a variety of sample sizes, system dimensions, weak and strong processes and DGPs.

We simulate VARMA(1,1) models expressed in Echelon form,

$$A_0 Y_t = A_1 Y_{t-1} + A_0 U_t + M_1 U_{t-1}, \tag{S.98}$$

using five alternative DGPs. DGPs I and II are designed such that all Kronecker indices are set to one, i.e. $\mathbf{p} = (p_1, p_2, ..., p_K)'$ with $p_i = 1$ for $i = 1, 2, ..., K$. These Kronecker indices yield $A_0 = I_K$, A_1 and M_1 as full matrices, and $2K^2$ free parameters. DGP I and II differ with respect to the eigenvalues associated with both AR and MA parameter matrices. The eigenvalues in DGP I are constant and equal to 0.5, while the eigenvalues driving II are potentially close to zero. DGPs III, IV and V impose the restrictions implied by Example 2, i.e. by setting the first k Kronecker indices to one and the remaining $K - k$ indices to zero. Precisely, the DGP III with $K = 3$ and presented in Tables 1 and S.4 is given by,

$$\begin{pmatrix} 1 & 0 & 0 \\ -0.6372 & 1 & 0 \\ -0.4372 & 0 & 1 \end{pmatrix} Y_t = \begin{pmatrix} 0.7724 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} Y_{t-1} + \begin{pmatrix} 1 & 0 & 0 \\ -0.6372 & 1 & 0 \\ -0.4372 & 0 & 1 \end{pmatrix} U_t + \begin{pmatrix} -0.4692 & 0.0380 & -0.0484 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} U_{t-1}. \tag{S.99}$$

Tables S.1, S.2 and S.3 contain the true vector of parameters used to simulate the different VARMA(1,1) specifications. It is important to highlight that the true parameters in DGPs III, IV and V are based on real data. This follows because they are the estimates obtained, on the first rolling window, from applying the IOLS estimator to the three key macroeconomic variables ($K = 3$), Dataset 1 ($K = 10$, $K = 20$ and $K = 40$), and to the entire dataset ($K = 52$).

The Monte Carlo results are organized as follows. Table S.4 reports results for the strong VARMA(1,1) model with $K = 3$ and DGPs I, II and III; Table S.5 and S.6 report results for strong and weak VARMA(1,1) models, respectively, with $K = 10$ and DGPs I, II and III; Table S.7 displays results for the strong VARMA(1,1) with $K = 10$, $K = 20$, $K = 40$ and $K = 52$ and DGPs III, IV and V; Table S.8 reports results for the strong and weak VARMA(1,1) models with $K = 10$ and $K = 20$, $T = 200$ and DGPs III, IV and V. For $K = 3$, we compare the IOLS estimator with the MLE, the two-stage of Hannan and Rissanen (1982) (HR), the three-stage of Hannan and Kavalieris (1984b) (HK), two-step of Dufour and Jouini (2014) (DJ2) and the three-stage multivariate version of Koreisha and Pukkila (1990) (KP), as discussed in Koreisha and Pukkila (2004) and Kascha (2012). The MLE estimator is implemented using Gauss 9.0 and the conditional log-likelihood function is maximized using the *Constrained Maximum Likelihood* (CML) library. We use the default specifications, so that the optimization procedure switches optimally over the different algorithms (BFGS (Broyden, Fletcher, Goldfarb, Shanno), DFP (Davidon, Fletcher, Powell), NEWTON (Newton-Raphson) and BHHH). For $K \geq 10$, we compare the IOLS algorithm to the HR, HK, DJ2, and KP estimators, because the MLE estimator is not feasible for high dimensional models.[2]

The first set of simulations, Table S.4, addresses the finite sample performance of the IOLS estimator in small sized, $K = 3$, strong VARMA models. The overall picture resembles the one in Table 1 in the article, i.e the MLE estimator is dominant in DGPs I and III, while the IOLS estimator outperforms all the competitors in DGP II. The good performance of the MLE estimator in DGPs I and III is justified by its asymptotic efficiency and corroborates results in previous studies (see Kascha (2012)). With regard to DGP II, we find that the IOLS estimator delivers the most accurate estimates. Recall that DGP II contains near-to-zero eigenvalues and hence poses further numerical challenges to estimation. We find that convergence rates for the IOLS estimator are low when $T \leq 150$, but increases monotonically with T. For $T = 400$, we find that IOLS marginally outperforms the MLE

[2]To assess the feasibility of the MLE estimator when applied to larger systems, we design Monte Carlo experiments considering VARMA models with eight and ten variables at different sample sizes: $T = 100$, $T = 150$, $T = 200$ and $T = 400$. We find that MLE is computationally infeasible for medium and large K. These results are available upon request.

estimator in terms of the MRRMSE measure, indicating that the MLE estimator improves its performance as T increases. This corroborates the fact that MLE is asymptotically efficient. Overall, we find that convergence rates from the IOLS estimator are higher (9% on average) in the strong VARMA(1,1) processes than in their weak counterpart.

We briefly discuss results for DGPs I, II and III and medium sized datasets ($K = 10$), as neither the strong nor the weak processes are included in the article. This set of simulations mimics the analyses with $K = 3$. We consider samples of 150, 200, 400 and 1000 observations. Table S.6 and S.5 display the results covering weak and strong processes, respectively. DGPs I and II yield 200 free parameters, while DGP III has 20 free parameters. Overall, we find that the best estimator on each model specification depends on the sample size and the eigenvalues associated with the parameter matrices. Considering DGP I, we find that the HK estimator is the one that produces the most accurate estimates for large sample sizes. This is the expected result, because the HK estimator is asymptotically efficient under Gaussian errors and DGP I does not pose numerical problems apart from the large number of free parameters. The second set of results in Tables S.6 and S.5 bring results for systems with near-to-zero eigenvalues. The DJ2 estimator delivers the strongest performance among all the competitors, while the IOLS estimator struggles in terms of convergence rates and performance. Nevertheless, convergence rate of the IOLS estimator increases monotonically with T, while its performance is equivalent to the HK estimator. Considering the DGP III, we now find that the HK estimator provides the most accurate estimates. The IOLS estimator improves drastically its relative performance, when compared to the previous set of simulations, positioning itself as the second best estimator in all sample sizes.

Tables S.7 S.8 bring the simulation results considering medium and high dimensional systems simulated with DGPs III, IV and V. Specifically, Table S.7 reports results for strong VARMA models with $T = 400$, while Table S.8 displays results for strong and weak processes with $T = 200$, $K = 10$ and $K = 20$. We do not present results for $T = 200$ and $K \geq 40$, because the maximum lag length in the VAR($\tilde{p}$ (first step in all estimators) is restricted to be at most $\tilde{p} = 4$. This follows from the limited number of degrees of freedom that is implied by high dimensional VAR models. Overall, we find that the IOLS estimator delivers an outstanding performance when $K \geq 40$ and $T = 400$ in all the three DGPs. In fact, the IOLS estimator outperforms all the competitors considering both MRRMSE and Share measures, delivering gains of up to 64% with respect to the HR benchmark

(MRRMSE measure). Convergence rates of the IOLS estimator are high (average of 98%) corroborating its strong performance. It worths highlighting that these are high dimensional models with up to 312 free parameters (DGP V and $K = 52$). When considering the results in Table S.8, we find that the IOLS estimator compares favourable to all the competitors for weak processes in DGPs IV and V (both $K = 10$ and $K = 20$). In fact, the HK only marginally beats the IOLS estimator for strong processes in DGPs IV and V. The main outcome, therefore, consists on showing that the IOLS estimator is particularly good at handling situations where K is large and T is small as well as the Kronecker indices are specified in the fashion of DGPs III, IV and V.

To wrap up, we conclude that the IOLS estimator outperforms the MLE and other linear estimators in a variety of cases, such as DGP II, high dimensional systems simulated with DGPs III, IV and V, and small sample sizes.

S. 5.1 Tables: Monte Carlo

Table S.1: Simulation Design, $K = 3$.

$K = 3$					
DGP I		DGP II		DGP II	
β_A	β_M	β_A	β_M	β_A	β_M
1.00	0.53	1.12	1.05	0.64	-0.47
0.23	-0.95	0.03	0.53	0.44	0.04
-0.05	-0.50	0.55	0.44	0.77	-0.05
-0.91	-0.01	0.26	0.15		
0.07	0.86	0.20	-0.15		
0.09	0.19	0.16	0.05		
0.75	0.02	-0.79	-0.44		
0.35	-0.74	0.54	-0.99		
0.43	0.11	-0.27	-0.22		

We display the free parameters used to simulate the models discussed in Tables 1 and S.4. In order to help the visualization, we partition the vector collecting the free parameters, β, in two vectors: β_A brings the free parameters associated with A_0 and A_1 parameter matrices, while β_M has the free parameters associated with the MA component. The free parameters are ordered so that $\beta = (\beta_A', \beta_M')'$ and $\text{vec}(B) = R\beta$ with $B = [(I_K - A_0), A_1, ..., A_p, M_1, ..., M_q]$ as in (8) and (9).

Table S.2: Simulation Design, $K = 10$ and $K = 20$.

$K=10$										$K=20$					
DGP I		DGP II		DGP III		DGP IV		DGP V		DGP III		DGP IV		DGP V	
β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M
0.05	0.03	0.02	0.01	-0.33	-0.06	0.57	-0.27	0.64	-0.28	0.33	-0.14	0.47	-0.37	0.66	-0.36
-0.13	0.07	-0.08	-0.09	-0.52	-0.03	0.58	-0.16	-0.91	-0.08	0.41	-0.05	0.61	-0.15	-0.86	-0.21
0.08	-0.11	-0.02	0.03	-0.21	0.09	-0.95	0.12	1.19	-0.20	0.54	-0.03	-0.92	-0.12	0.52	-0.21
0.05	-0.19	-0.15	0.02	0.00	-0.07	1.04	-0.02	-0.15	0.11	-0.79	0.01	0.42	-0.05	1.32	-0.12
0.04	0.14	0.10	0.10	-0.35	0.16	-0.12	0.01	-0.13	0.04	0.28	0.01	1.53	-0.01	0.90	0.03
-0.02	0.10	-0.10	0.15	-0.50	0.06	-0.11	0.05	-0.09	0.19	1.47	0.02	0.98	-0.04	1.01	0.17
0.01	-0.14	0.16	0.21	-0.70	0.06	-0.03	0.05	-0.16	0.05	0.84	0.20	0.92	0.09	0.23	-0.03
0.14	-0.21	0.02	0.12	-0.41	0.07	0.21	0.02	-0.03	-0.34	0.32	0.14	0.11	0.01	-0.06	-0.39
0.01	-0.06	0.12	0.10	0.46	0.08	-0.19	-0.04	0.30	0.05	0.06	0.07	-0.07	-0.06	0.68	-0.04
0.00	-0.04	-0.03	-0.09			-0.07	-0.13	-0.65	0.04	0.00	0.09	0.64	0.01	0.45	0.09
0.68	0.47	0.14	0.55			0.26	-0.11	0.96	0.09	0.57	0.00	0.52	-0.04	0.28	0.14
0.25	-0.01	0.15	0.05			-0.82	0.11	0.89	0.01	0.40	0.03	0.29	0.03	-0.05	0.00
0.30	0.00	0.02	-0.03			0.99	0.02	0.47	-0.04	0.24	-0.08	-0.04	-0.01	0.01	-0.06
0.00	0.01	0.02	0.00			0.91	0.10	-0.18	0.07	0.66	0.01	0.03	0.24	0.06	0.03
-0.24	0.03	0.02	0.09			0.53	0.00	-0.22	-0.13	0.39	-0.05	0.02	0.12	-0.07	0.00
0.31	-0.02	-0.14	0.18			0.80	0.06	-0.18	-0.12	-0.23	0.13	-0.05	0.14	-0.37	-0.04
0.06	0.01	-0.08	0.44			0.38	0.30	-0.90	0.01	0.69	0.00	-0.34	0.12	-0.05	0.00
0.00	0.07	0.14	-0.11			0.01	0.05	0.17	0.08	0.28	0.19	0.20	0.10	-0.05	0.02
0.13	0.09	-0.21	-0.02			0.01	0.18	0.10	-0.01	0.85	-0.10	-0.08	-0.13	0.20	-0.01
0.17	-0.10	0.04	0.00					0.27	-0.02			-0.09	0.08	-0.15	0.04
-0.18	-0.02	0.03	0.18					0.83	0.13			0.19	0.30	-0.22	0.42
0.37	0.45	0.15	0.28					0.39	0.02			-0.20	0.01	-0.17	0.14
-0.29	-0.04	-0.42	-0.06					0.07	-0.35			-0.10	-0.01	-0.64	0.21
-0.01	0.01	-0.11	0.27					0.03	0.05			-0.22	0.04	0.00	0.14
0.09	0.06	0.07	0.01					0.00	0.30			-0.81	0.01	0.00	0.11
-0.23	-0.02	0.07	0.10					0.23	0.12			-0.09	-0.09	-0.07	-0.09
-0.07	0.03	-0.20	0.16					-0.19	0.06			0.05	0.08	-0.18	0.09
-0.08	0.02	0.06	0.09					-0.02	0.18			-0.11	0.01	-0.04	-0.13
-0.04	0.06	-0.08	0.13					0.59	0.06			-0.15	0.05	0.90	-0.06
0.06	0.12	-0.08	-0.06									-0.07	-0.07	0.45	0.08
0.16	0.01	0.13	-0.22									0.91	0.20	-0.24	0.30
0.31	0.12	-0.15	-0.19									0.48	0.16	0.95	0.01
0.77	0.42	0.27	0.16									-0.32	0.11	0.76	0.00
-0.03	-0.23	0.24	0.04									0.96	0.00	-0.17	-0.01
-0.26	0.04	0.22	0.18									0.84	0.13	-0.20	0.00
0.32	0.17	-0.18	-0.13									0.85	0.19	-0.15	0.02
0.07	-0.04	0.29	-0.13									0.43	0.14	-0.31	0.00
-0.06	-0.18	0.05	0.06									-0.02	-0.13	0.74	-0.20
0.09	-0.15	0.04	-0.14									0.16	0.08	-0.03	-0.07
-0.06	0.05	-0.15	-0.29											-0.72	0.09
0.17	-0.01	0.01	-0.49											-0.47	0.08
0.16	-0.04	-0.10	-0.07											0.17	-0.01
0.28	0.02	-0.03	0.16											-0.18	0.05
0.51	0.56	0.09	0.14											0.22	0.05
-0.18	-0.08	0.08	0.20											-0.08	-0.06
0.23	-0.01	0.02	-0.27											0.03	0.20
0.08	0.13	0.00	-0.07											0.11	0.10
0.07	0.05	-0.06	-0.07											-0.32	0.13
0.05	-0.05	0.21	-0.11											0.08	0.11
0.03	0.02	0.03	-0.05											0.30	0.06
0.01	-0.02	0.01	-0.08											0.82	-0.01
-0.04	-0.02	0.20	-0.01											0.42	0.13
0.03	0.06	0.13	0.14											0.15	0.07
0.03	0.08	0.03	0.06											0.04	0.18
0.42	0.45	0.17	-0.01											0.15	0.14
0.00	-0.05	-0.07	-0.13											0.20	-0.35
-0.01	0.04	0.04	-0.13											-0.16	-0.09
0.05	0.11	-0.03	0.12											0.03	0.09
0.06	0.06	0.08	-0.16											0.59	0.06

Continued on next page

Continued from previous page

K = 10										K = 20					
DGP I		DGP II		DGP III		DGP IV		DGP V		DGP III		DGP IV		DGP V	
β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M
0.11	0.17	0.06	0.07												
0.11	0.03	-0.31	0.27												
0.18	0.31	0.07	0.03												
0.18	-0.08	-0.46	0.00												
-0.01	-0.44	0.02	0.18												
-0.25	0.15	-0.27	0.03												
0.68	0.76	0.38	0.51												
0.06	-0.36	-0.39	0.48												
0.00	-0.34	0.09	-0.18												
0.07	-0.14	-0.16	0.12												
0.00	-0.03	0.14	-0.04												
-0.01	0.08	-0.07	0.08												
0.05	-0.06	0.16	0.01												
0.01	0.04	0.09	0.02												
0.00	0.15	0.23	0.06												
0.00	-0.06	0.01	-0.10												
0.05	-0.07	0.05	0.08												
0.47	0.45	0.10	0.29												
-0.06	-0.07	0.05	0.11												
0.06	-0.08	-0.08	-0.08												
-0.04	-0.04	0.19	0.18												
0.02	0.04	-0.05	0.05												
0.15	-0.21	0.29	0.13												
0.04	0.03	-0.38	0.00												
-0.04	0.28	0.07	-0.07												
0.05	-0.16	-0.28	0.16												
0.12	-0.12	0.36	-0.01												
0.01	0.26	-0.43	0.20												
0.39	0.57	0.09	0.25												
-0.02	-0.11	-0.01	-0.08												
0.03	0.05	0.11	-0.07												
0.04	0.00	-0.28	-0.16												
0.05	-0.13	0.01	-0.09												
0.02	0.05	-0.06	-0.09												
-0.05	0.20	0.13	0.02												
-0.04	-0.17	-0.01	0.05												
-0.01	-0.08	0.00	0.02												
0.08	0.24	-0.14	0.09												
0.03	0.12	0.14	-0.13												
0.36	0.42	0.14	0.00												

We display the free parameters used to simulate the models discussed in Tables 2, S.5, S.6, S.7 and S.8. In order to help the visualization, we partition the vector collecting the free parameters, β, in two vectors: β_A brings the free parameters associated with A_0 and A_1 parameter matrices, while β_M has the free parameters associated with the MA component. The free parameters are ordered so that $\beta = \left(\beta_A', \beta_M'\right)'$ and $\text{vec}\,(B) = R\beta$ with $B = [(I_K - A_0), A_1, ..., A_p, M_1, ..., M_q]$ as in (8) and (9).

Table S.3: Simulation Design, $K = 20$, $K = 40$ and $K = 52$.

$K=40$						$K=52$					
DGP III		DGP IV		DGP V		DGP III		DGP IV		DGP V	
β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M
0.63	-0.57	0.08	-0.47	0.41	-0.47	0.59	-0.44	0.05	-0.35	0.34	-0.35
0.21	-0.05	0.40	-0.16	0.59	-0.15	0.18	0.05	0.34	-0.12	0.48	-0.12
0.39	0.01	0.58	-0.21	-0.79	0.00	0.35	0.01	0.48	-0.11	0.32	0.00
0.53	-0.06	-0.81	-0.31	0.39	-0.26	0.48	-0.04	0.29	-0.30	0.56	-0.15
-0.75	0.08	0.36	0.02	1.02	-0.30	0.26	0.00	0.56	0.01	0.31	-0.28
0.29	0.07	1.09	0.00	0.76	-0.11	0.54	0.10	0.30	-0.02	1.08	-0.14
1.22	0.13	0.78	-0.07	0.66	0.22	0.24	0.06	1.12	-0.06	1.02	0.16
0.74	0.16	0.63	0.07	0.04	-0.07	1.19	-0.17	1.01	0.02	0.61	-0.08
0.28	0.13	0.02	0.08	-0.08	-0.38	0.99	0.22	0.62	0.00	0.71	-0.41
0.01	0.12	-0.09	-0.10	0.47	-0.07	0.61	-0.32	0.73	0.10	0.43	-0.07
-0.01	0.10	0.45	0.05	0.30	0.08	0.68	0.08	0.45	0.09	-0.77	0.02
0.44	-0.02	0.32	-0.04	0.21	0.09	0.48	0.05	-0.79	0.00	0.84	0.04
0.29	0.01	0.21	0.14	-0.06	0.07	-0.73	0.05	0.86	0.06	0.47	0.01
0.19	-0.11	-0.05	0.01	-0.02	-0.10	0.88	0.05	0.46	-0.11	0.94	0.10
0.55	0.02	-0.01	0.18	0.09	0.10	0.50	0.20	0.99	-0.17	1.08	0.05
0.31	0.00	0.06	0.15	-0.05	0.05	1.11	-0.03	1.12	0.01	0.45	0.08
-0.14	0.10	-0.03	0.11	-0.32	-0.04	1.18	-0.70	0.50	0.24	0.30	0.00
0.59	-0.11	-0.31	0.09	1.19	-0.01	0.63	0.71	0.36	-0.11	0.06	0.10
0.20	0.07	1.24	0.16	1.01	0.13	0.56	0.26	0.04	-0.32	-0.04	0.06
1.31	-0.05	1.00	-0.11	1.20	0.02	0.01	-0.04	-0.06	0.05	0.06	-0.11
0.98	-0.05	1.25	0.10	0.19	0.11	-0.19	0.10	0.04	0.08	0.59	0.10
1.30	-0.21	0.21	0.29	0.01	0.30	-0.02	0.00	0.57	0.16	0.06	-0.17
0.27	0.45	0.02	-0.01	-0.21	0.15	0.26	-0.08	0.05	0.05	-0.01	0.01
0.52	0.04	-0.23	-0.04	-0.10	0.55	0.16	0.12	-0.03	-0.05	-0.06	-0.17
-0.01	-0.02	-0.07	0.02	-0.06	0.11	0.07	-0.03	-0.06	0.04	0.20	0.15
-0.11	0.04	-0.02	0.02	0.88	0.09	-0.32	0.04	0.21	-0.03	-0.08	-0.12
0.11	0.00	0.91	-0.12	-0.03	-0.09	0.25	-0.06	-0.09	0.04	0.35	-0.20
0.93	-0.30	-0.07	0.06	0.35	0.15	-0.03	0.04	0.33	-0.16	0.25	-0.33
-0.19	0.20	0.33	0.03	0.30	-0.12	0.36	-0.01	0.26	0.20	0.05	0.05
0.24	-0.05	0.40	0.08	0.09	-0.02	0.25	-0.01	0.03	0.06	0.03	-0.01
0.62	-0.21	0.07	-0.05	-0.11	0.11	-0.12	-0.11	0.02	-0.03	-0.04	0.09
0.20	-0.19	-0.12	0.01	0.05	0.29	-0.12	-0.94	-0.03	-0.05	0.03	0.16
-0.36	-0.02	0.01	0.15	-0.13	-0.03	0.18	0.98	0.00	-0.75	-0.05	0.05
-0.16	-0.05	-0.12	-0.02	0.14	-0.01	-0.11	0.14	-0.04	0.20	-0.06	0.04
0.11	0.09	0.20	-0.10	-0.02	-0.04	0.54	0.07	-0.06	0.70	-0.02	-0.05
0.32	0.04	0.01	0.22	-0.07	-0.04	0.51	-0.05	-0.02	-0.06	-0.01	-0.13
0.12	0.03	-0.10	0.07	0.02	0.01	0.51	-0.18	0.00	0.28	-0.02	0.05
0.09	0.20	-0.01	0.08	0.00	0.02	0.48	0.20	-0.02	-0.02	-0.05	-0.03
0.12	-0.09	0.24	-0.11	-0.05	-0.17	0.37	0.06	-0.05	-0.05	-0.15	0.01
0.71	0.06	-0.03	0.07	0.15	-0.11	0.28	-0.36	-0.14	0.02	-0.31	0.03
		-0.08	-0.04	-0.06	0.06	0.26	0.06	-0.30	0.10	-0.09	-0.16
		0.12	-0.01	0.00	0.06	0.18	0.44	-0.11	0.30	-0.01	-0.01
		-0.11	-0.21	-0.09	0.02	0.08	-0.25	-0.02	-0.01	-0.21	0.20
		0.14	0.05	-0.42	0.08	0.10	-0.05	-0.23	0.06	-0.08	0.06
		-0.10	0.43	0.07	-0.01	-0.01	0.05	-0.06	-0.09	-0.09	-0.01
		-0.58	-0.04	0.07	-0.05	-0.11	0.02	-0.09	-0.02	-0.05	-0.03
		-0.02	0.05	0.01	0.01	0.10	0.02	-0.02	0.15	-0.02	-0.05
		0.11	0.03	-0.11	0.20	0.10	0.06	0.01	-0.13	0.13	0.02
		-0.01	-0.01	0.00	0.13	0.11	-0.33	0.18	-0.04	0.17	-0.50
		-0.05	0.38	0.92	-0.02	0.31	0.22	0.18	0.00	0.05	0.20
		-0.02	0.02	0.49	0.13	0.19	0.00	0.24	0.02	0.02	0.93
		0.92	0.09	-0.22	-0.10	0.66	0.03	0.01	-0.08	0.01	0.62
		0.52	0.00	0.97	0.22			0.00	-0.05	-0.02	-0.06
		-0.30	-0.02	0.78	0.01			-0.04	-0.04	-0.06	-0.39
		0.97	-0.31	-0.02	0.07			-0.05	0.05	0.02	0.24
		0.84	-0.20	0.00	0.08			-0.10	0.03	0.00	-0.02
		0.04	0.20	-0.03	-0.38			0.08	-0.01	-0.03	-0.16
		-0.03	0.06	0.03	-0.10			-0.03	-0.03	-0.10	-0.06
		0.03	-0.07	0.79	0.07			-0.04	-0.01	0.01	0.03
		0.07	0.07	0.33	0.04			-0.11	-0.01	0.13	0.01

Continued on next page

Continued from previous page

K = 40						K = 52					
DGP III		DGP IV		DGP V		DGP III		DGP IV		DGP V	
β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M
		0.79	-0.20	-0.08	-0.08			0.05	-0.13	-0.01	0.11
		0.34	0.00	0.11	-0.01			0.10	0.07	0.07	0.30
		-0.06	-0.22	-0.06	-0.07			0.06	-0.82	0.02	-0.02
		0.21	0.03	-0.03	-0.23			0.09	-0.33	0.02	-0.04
		0.01	-0.04	-0.13	0.05			0.16	0.82	-0.03	0.06
		-0.19	-0.03	0.11	-0.03			0.07	0.38	0.10	-0.08
		-0.17	-0.05	0.24	0.39			0.19	0.08	0.01	-0.08
		0.34	-0.06	-0.37	-0.04			0.32	-0.22	-0.06	-0.01
		0.20	0.08	-0.14	-0.26			-0.07	0.05	-0.01	0.01
		-0.37	-0.09	0.34	0.03			-0.20	-0.12	-0.41	0.14
		-0.24	0.02	0.08	0.03			-0.09	-0.05	0.21	-0.13
		0.34	-0.05	0.10	0.04			-0.54	-0.04	0.26	-0.04
		0.20	0.04	0.29	-0.01			0.18	-0.25	-0.42	-0.04
		0.18	0.02	0.19	0.38			0.19	0.29	0.00	0.00
		0.28	0.18	-0.14	-0.09			-0.43	0.21	0.03	0.15
		0.17	0.17	-0.14	0.02			0.04	-0.63	0.09	0.01
		0.62	-0.06	-0.14	0.09			0.09	0.11	-0.07	-0.08
		0.20	-0.19	-0.25	0.11			0.07	0.87	-0.17	-0.09
		0.11	0.06	0.66	0.00			-0.01	-0.43	-0.14	-0.05
		0.33	-0.01	-0.01	-0.02			-0.27	-0.31	0.28	-0.04
				-0.59	-0.09			-0.24	0.16	-0.07	0.08
				-0.36	-0.31			0.37	0.05	0.96	0.04
				0.13	-0.20			-0.19	0.43	0.93	0.03
				-0.13	0.16			0.98	0.17	0.87	0.06
				0.23	0.20			0.93	-0.29	0.79	0.00
				-0.05	0.06			0.88	0.02	0.63	-0.03
				0.02	0.02			0.80	-0.04	0.51	-0.05
				0.09	-0.08			0.66	-0.17	0.63	-0.02
				-0.34	0.07			0.54	0.07	0.76	-0.01
				0.06	-0.07			0.68	-0.01	0.30	-0.17
				0.22	-0.20			0.83	0.01	0.18	-0.13
				0.33	0.00			0.29	0.09	0.33	0.07
				-0.10	-0.14			0.17	0.02	-0.11	0.07
				0.28	-0.20			0.33	-0.01	0.31	-0.95
				0.15	0.03			-0.08	0.05	0.10	-0.34
				0.03	0.03			0.32	-0.03	0.09	-1.25
				-0.04	-0.04			0.20	-0.33	0.09	0.96
				0.14	-0.03			0.17	-0.21	0.04	0.39
				0.40	0.11			0.20	0.22	-0.13	1.22
				0.33	-0.04			0.02	0.18	-0.06	0.10
				-0.64	-0.06			0.58	0.00	-0.24	-0.22
				-0.15	0.05			0.16	0.02	-0.12	-0.01
				1.00	0.07			0.11	0.03	-0.15	0.05
				-0.19	-0.09			0.38	0.07	0.33	-0.12
				-0.03	0.10					-0.10	0.13
				-0.46	0.03					0.01	-0.04
				0.04	-0.05					0.01	-0.04
				0.55	0.06					0.17	-0.37
				0.34	0.04					-0.15	-0.26
				-0.10	0.02					0.30	0.28
				-0.10	-0.03					0.04	0.07
				0.62	0.18					0.65	0.23
				0.19	0.17					0.28	-0.63
				0.02	-0.01					0.88	0.42
				0.14	-0.06					0.94	0.13
				0.32	-0.19					-0.31	0.86
				0.30	-0.24					-0.59	-0.45
				-0.14	0.06					-0.34	-0.47
				0.06	-0.01					-0.52	-0.32
				0.53	0.02					-0.14	0.24

Continued on next page

Continued from previous page

$K = 40$						$K = 52$					
DGP III		DGP IV		DGP V		DGP III		DGP IV		DGP V	
β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M	β_A	β_M
										-0.31	0.17
										-0.04	0.06
										0.17	0.00
										0.17	0.42
										-0.11	0.17
										0.22	-0.13
										-0.37	-0.27
										-0.36	0.02
										0.38	0.08
										-0.49	-0.03
										0.08	-0.17
										0.01	-0.29
										0.05	0.06
										0.06	-0.01
										0.10	0.03
										0.15	0.00
										0.17	0.10
										0.26	0.14
										-0.07	0.03
										-0.06	-0.01
										-0.02	-0.12
										0.14	0.06
										0.05	-0.03
										0.41	0.08
										0.36	-0.35
										0.52	-0.21
										-0.02	0.16
										0.58	0.22
										0.16	0.18
										0.02	-0.02
										0.14	0.00
										0.36	0.01
										0.30	-0.02
										-0.10	0.02
										0.06	0.07
										0.53	-0.01

We display the free parameters used to simulate the models discussed in Tables 2 and S.7. In order to help the visualization, we partition the vector collecting the free parameters, β, in two vectors: β_A brings the free parameters associated with A_0 and A_1 parameter matrices, while β_M has the free parameters associated with the MA component. The free parameters are ordered so that $\beta = (\beta_A', \beta_M')'$ and $\text{vec}(B) = R\beta$ with $B = [(I_K - A_0), A_1, ..., A_p, M_1, ..., M_q]$ as in (8) and (9).

Table S.4: Monte Carlo - Strong VARMA(1,1) models: Small Sized Systems, $K = 3$.

	DGP I						DGP II						DGP III					
	T=50, $n = 18$						T=50, $n = 18$						T=50, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	0.86	**0.78**	1.00	0.94	0.84	1.00	1.03	1.03	1.20	**0.97**	1.29	1.00	0.59	0.69	0.69	1.00	**0.53**
Share (%)	0%	**39%**	22%	0%	17%	22%	0%	28%	11%	0%	**39%**	22%	0%	33%	0%	0%	0%	**67%**
Convergence(%)	100%	62%	90%	92%	93%	60%	100%	17%	92%	73%	95%	18%	100%	69%	100%	89%	99%	78%
	T=100, $n = 18$						T=100, $n = 18$						T=100, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	0.98	**0.88**	0.98	0.99	0.90	1.00	1.00	1.04	1.22	**0.99**	1.31	1.00	0.74	0.83	0.69	1.26	**0.63**
Share (%)	11%	17%	17%	0%	6%	**50%**	0%	**50%**	6%	0%	33%	11%	0%	33%	0%	0%	0%	**67%**
Convergence(%)	100%	81%	97%	99%	98%	92%	100%	32%	97%	79%	99%	59%	99%	74%	100%	93%	100%	97%
	T=150, $n = 18$						T=150, $n = 18$						T=150, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.08	1.13	1.02	0.99	**0.97**	1.00	**0.93**	1.01	1.07	1.00	1.13	1.00	0.86	1.68	0.78	1.45	**0.74**
Share (%)	0%	6%	6%	11%	22%	**56%**	6%	**39%**	33%	0%	0%	22%	0%	33%	0%	0%	0%	**67%**
Convergence(%)	100%	88%	96%	99%	99%	95%	100%	45%	97%	85%	100%	72%	100%	79%	100%	98%	100%	99%
	T=200, $n = 18$						T=200, $n = 18$						T=200, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.09	1.11	1.02	1.00	**0.98**	1.00	**0.90**	1.03	1.00	1.00	1.01	1.00	0.88	1.70	0.78	1.40	**0.76**
Share (%)	11%	6%	6%	6%	22%	**50%**	6%	**39%**	33%	0%	0%	22%	0%	33%	0%	0%	0%	**67%**
Convergence(%)	100%	91%	99%	100%	100%	97%	100%	54%	98%	91%	100%	77%	100%	82%	100%	99%	100%	100%
	T=400, $n = 18$						T=400, $n = 18$						T=400, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.10	1.15	1.00	1.00	**0.98**	1.00	**0.78**	1.11	0.81	1.00	0.79	1.00	0.92	1.92	0.78	1.27	**0.78**
Share (%)	22%	6%	0%	17%	11%	**44%**	0%	**39%**	22%	0%	0%	**39%**	0%	0%	0%	33%	0%	**67%**
Convergence(%)	100%	96%	100%	100%	100%	99%	100%	74%	100%	96%	100%	87%	100%	86%	100%	100%	100%	100%

We report results from strong VARMA(1,1) models simulated with different Kronecker indices. The first column reports results from DGP I, while the second and third columns display results from strong VARMA(1,1) models simulated with DGPs II and III, respectively. Recall that DGP I and II set all the Kronecker indices to one, $\mathbf{p} = (1, 1, 1)'$, while DGP III sets $\mathbf{p} = (1, 0, 0)'$. DGPs I and II differ with respect to the eigenvalues driving the AR and MA parameter matrices. DGP I has all the eigenvalues associated with both the AR and MA parameter matrices set to 0.5, while DGP II has eigenvalues associated with the AR and MA components given by $(0.80, 0.20, 0.05)'$ and $(0.90, -0.02, -0.20)'$, respectively. The true vector of parameters in DGP III contains the estimates obtained by fitting a VARMA(1,1) model on the first rolling window of a dataset comprising only the three key macroeconomic variables studied in Section 5. Table S.1 displays the true values used to simulate DGPs I, II and III. n accounts for the number of free parameters in the model. MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share % is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage of Hannan and Kavalieris (1984b); KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012); and MLE accounts for the maximum likelihood estimator. The number of replications is set to 5000.

Table S.5: Monte Carlo - Strong VARMA(1,1) models: Medium Sized Systems, $K = 10$.

	DGP I					DGP II					DGP III				
	T=150, $n = 200$					T=150, $n = 200$					T=150, $n = 20$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.94	1.02	1.10	**0.93**	1.00	1.29	1.00	1.43	**0.93**	1.00	0.69	1.21	**0.68**	1.01
Share (%)	1%	**51%**	23%	2%	24%	1%	14%	9%	2%	**76%**	0%	40%	0%	**55%**	5%
Convergence(%)	100%	41%	81%	96%	95%	100%	5%	97%	90%	100%	100%	96%	99%	100%	98%
	T=200, $n = 200$					T=200, $n = 200$					T=200, $n = 20$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.99	1.04	**0.97**	1.01	1.00	0.99	**0.76**	1.07	0.92	1.00	0.98	1.76	**0.83**	1.14
Share (%)	2%	19%	**44%**	23%	13%	2%	4%	**74%**	1%	20%	20%	15%	5%	**60%**	0%
Convergence(%)	100%	62%	86%	98%	99%	100%	16%	100%	96%	100%	100%	98%	100%	100%	98%
	T=400, $n = 200$					T=400, $n = 200$					T=400, $n = 20$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	1.00	1.13	**0.93**	1.06	1.00	1.09	**0.82**	1.12	0.98	1.00	0.99	2.20	**0.79**	1.20
Share (%)	1%	3%	34%	**55%**	8%	7%	2%	**74%**	2%	17%	20%	10%	5%	**65%**	0%
Convergence(%)	100%	87%	91%	99%	100%	100%	56%	100%	99%	100%	100%	99%	100%	100%	99%
	T=1000, $n = 200$					T=1000, $n = 200$					T=1000, $n = 20$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	1.02	1.06	**0.92**	1.05	1.00	1.12	**0.92**	1.10	1.01	1.00	0.99	2.20	**0.76**	1.34
Share (%)	2%	0%	23%	**72%**	4%	4%	1%	**68%**	9%	20%	0%	15%	5%	**80%**	0%
Convergence(%)	100%	95%	98%	100%	100%	100%	87%	100%	100%	100%	100%	100%	100%	100%	100%

We report results from strong VARMA(1,1) models simulated with different Kronecker indices. The first column reports results from DGP I, while the second and third columns display results from strong VARMA(1,1) models simulated with DGPs II and III, respectively. Recall that DGP I and II set all the Kronecker indices to one, $\mathbf{p} = (1, 1, 1, ..., 1)'$, while DGP III sets $\mathbf{p} = (1, 0, 0, 0, ..., 0)'$. DGPs I and II differ with respect to the eigenvalues driving the AR and MA parameter matrices. DGP I has all the eigenvalues associated with both the AR and MA parameter matrices set to 0.5, while DGP II has eigenvalues associated with the AR and MA components given by $(0.03, 0.05, -0.30, -0.30, 0.20, 0.20, 0.50, 0.50, 0.50, 0.50)'$ and $(-0.20, -0.02, 0.03, 0.03, 0.05, 0.40, 0.50, 0.50, 0.50, 0.50)'$, respectively. The true vector of parameters in DGP III contains the estimates obtained by fitting a VARMA(1,1) model on the first rolling window of Dataset 1 with $K = 10$. Table S.2 displays the true values used to simulate DGPs I, II and III. n accounts for the number of free parameters in the model. MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share % is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage of Hannan and Kavalieris (1984b); and KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012). The number of replications is set to 5000.

Table S.6: Monte Carlo - Weak VARMA(1,1) models: Medium Sized Systems, $K = 10$.

	DGP I					DGP II					DGP III				
	T=150, $n = 200$					T=150, $n = 200$					T=150, $n = 20$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.98	1.04	1.14	**0.94**	1.00	1.18	0.98	1.31	**0.94**	1.00	0.69	1.20	**0.66**	1.01
Share (%)	1%	**48%**	33%	2%	17%	2%	16%	40%	1%	**43%**	0%	40%	0%	**55%**	5%
Convergence(%)	100%	17%	54%	92%	90%	100%	4%	77%	89%	98%	100%	97%	97%	100%	98%
	T=200, $n = 200$					T=200, $n = 200$					T=200, $n = 20$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	1.02	1.02	1.03	**0.98**	1.00	1.10	**0.86**	1.17	0.95	1.00	0.99	1.71	**0.80**	1.14
Share (%)	4%	23%	**39%**	13%	22%	3%	2%	**68%**	3%	26%	15%	15%	5%	**65%**	0%
Convergence(%)	100%	32%	66%	96%	97%	100%	13%	95%	93%	100%	100%	98%	100%	100%	96%
	T=400, $n = 200$					T=400, $n = 200$					T=400, $n = 20$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	1.03	1.08	**0.98**	1.03	1.00	1.12	**0.84**	1.15	0.98	1.00	0.99	2.00	**0.76**	1.17
Share (%)	7%	5%	33%	**39%**	17%	7%	2%	**77%**	2%	13%	0%	25%	5%	**70%**	0%
Convergence(%)	100%	67%	77%	99%	100%	100%	45%	99%	98%	100%	100%	98%	100%	100%	98%
	T=1000, $n = 200$					T=1000, $n = 200$					T=1000, $n = 20$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	1.05	1.01	**0.97**	1.04	1.00	1.15	**0.90**	1.14	1.00	1.00	0.98	1.86	**0.73**	1.26
Share (%)	7%	2%	37%	**44%**	11%	5%	1%	**75%**	5%	16%	0%	25%	5%	**70%**	0%
Convergence(%)	100%	88%	91%	100%	100%	100%	78%	100%	100%	100%	100%	100%	100%	100%	99%

We report results from weak VARMA(1,1) models simulated with different Kronecker indices. The first column reports results from DGP I, while the second and third columns display results from weak VARMA(1,1) models simulated with DGPs II and III, respectively. Recall that DGP I and II set all the Kronecker indices to one, $\mathbf{p} = (1, 1, 1, ..., 1)'$, while DGP III sets $\mathbf{p} = (1, 0, 0, 0, ..., 0)'$. DGPs I and II differ with respect to the eigenvalues driving the AR and MA parameter matrices. DGP I has all the eigenvalues associated with both the AR and MA parameter matrices set to 0.5, while DGP II has eigenvalues associated with the AR and MA components given by $(0.03, 0.05, -0.30, -0.30, 0.20, 0.20, 0.50, 0.50, 0.50, 0.50)'$ and $(-0.20, -0.02, 0.03, 0.03, 0.05, 0.40, 0.50, 0.50, 0.50, 0.50)'$, respectively. The true vector of parameters in DGP III contains the estimates obtained by fitting a VARMA(1,1) model on the first rolling window of Dataset 1 with $K = 10$. Table S.2 displays the true values used to simulate DGPs I, II and III. n accounts for the number of free parameters in the model. Weak innovations are generated using the procedure described in Romano and Thombs (1996). MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share % is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage of Hannan and Kavalieris (1984b); and KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012). The number of replications is set to 5000.

Table S.7: Monte Carlo - Strong VARMA(1,1) models: Medium and Large Sized Systems, $K = 10$, $K = 20$, $K = 40$ and $K = 52$, with $T = 400$.

	DGP III					DGP IV					DGP V				
	K=10, $n = 20$					K=10, $n = 40$					K=10, $n = 60$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.99	2.20	**0.79**	1.20	1.00	0.94	1.94	**0.77**	1.34	1.00	0.94	1.59	**0.81**	1.25
Share (%)	20%	10%	5%	**65%**	0%	0%	23%	5%	**73%**	0%	0%	13%	13%	**72%**	2%
Convergence(%)	100%	99%	100%	100%	99%	100%	96%	100%	100%	97%	100%	85%	100%	100%	96%
	K=20, $n = 40$					K=20, $n = 80$					K=20, $n = 120$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.99	2.43	**0.70**	1.36	1.00	0.99	1.90	**0.78**	1.33	1.00	0.98	1.75	**0.80**	1.39
Share (%)	25%	8%	3%	**65%**	0%	18%	19%	3%	**61%**	0%	13%	14%	5%	**66%**	2%
Convergence(%)	100%	100%	100%	100%	89%	100%	99%	100%	100%	85%	100%	97%	100%	100%	85%
	K=40, $n = 80$					K=40, $n = 160$					K=40, $n = 240$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	**0.36**	1.22	0.45	1.02	1.00	**0.44**	1.26	0.59	1.22	1.00	**0.46**	1.31	0.61	1.39
Share (%)	4%	**51%**	0%	45%	0%	9%	**58%**	0%	34%	0%	10%	**61%**	0%	29%	0%
Convergence(%)	100%	100%	96%	100%	70%	100%	100%	97%	100%	63%	100%	100%	93%	100%	48%
	K=52, $n = 104$					K=52, $n = 208$					K=52, $n = 312$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	**0.43**	1.92	0.43	2.30	1.00	**0.44**	1.67	0.52	2.07	1.00	**0.50**	1.88	0.57	2.80
Share (%)	4%	**48%**	0%	**48%**	0%	6%	**53%**	0%	41%	0%	9%	**56%**	0%	35%	0%
Convergence(%)	100%	100%	88%	98%	37%	100%	100%	84%	98%	26%	100%	100%	70%	96%	9%

We report results from strong VARMA(1,1) models simulated with different Kronecker indices, system sizes and fixed sample size set as $T = 400$. The first column reports results from DGP III, while the second and third columns of results display results from weak VARMA(1,1) models simulated from DGPs IV and V, respectively. Recall that DGP III sets the first Kronecker index to one and all the remaining indices to zero, $\mathbf{p} = (1, 0, 0, 0, ..., 0)'$; DGP IV sets the first two Kronecker indices to one and the remaining $K - 2$ indices to zero, $\mathbf{p} = (1, 1, 0, 0, ..., 0)'$; and DGP V sets the first three Kronecker indices to one and the remaining $K - 3$ indices to zero, $\mathbf{p} = (1, 1, 1, 0, ..., 0)'$. The true vectors of parameters in DGP III, IV and V contain the estimates obtained by fitting VARMA(1,1) models on the first rolling window of Dataset 1 in their respective system dimensions. Tables S.2 and S.3 display the true values used to simulate DGPs III, IV and V. n accounts for the number of free parameters in the model. MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share % is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage of Hannan and Kavalieris (1984b); and KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012). The number of replications is set to 1000.

Table S.8: Monte Carlo - Strong and Weak VARMA(1,1) models: Medium Sized Systems, $K = 10$ and $K = 20$, with $T = 200$.

	DGP III, Strong					DGP IV, Strong					DGP V, Strong				
	K=10, $n = 20$					K=10, $n = 40$					K=10, $n = 60$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.98	1.76	**0.83**	1.14	1.00	0.75	1.27	**0.73**	1.12	1.00	**0.78**	1.16	0.79	1.07
Share (%)	20%	15%	5%	**60%**	0%	0%	43%	5%	**45%**	8%	5%	**43%**	10%	35%	7%
Convergence(%)	100%	98%	100%	100%	98%	100%	92%	99%	99%	97%	99%	73%	99%	98%	97%
	K=20, $n = 40$					K=20, $n = 80$					K=20, $n = 120$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.75	1.66	**0.64**	1.45	1.00	0.78	1.39	**0.73**	1.23	1.00	0.77	1.30	**0.74**	1.22
Share (%)	5%	43%	0%	**53%**	0%	5%	40%	3%	**53%**	0%	8%	38%	3%	**48%**	3%
Convergence(%)	100%	100%	100%	100%	88%	100%	97%	100%	100%	84%	100%	94%	100%	100%	75%

	DGP III, Weak					DGP IV, Weak					DGP V, Weak				
	K=10, $n = 20$					K=10, $n = 40$					K=10, $n = 60$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.99	1.71	**0.80**	1.14	1.00	**0.73**	1.21	0.79	1.08	1.00	**0.77**	1.17	0.85	1.08
Share (%)	15%	15%	5%	**65%**	0%	0%	**68%**	8%	13%	13%	7%	**63%**	8%	17%	5%
Convergence(%)	100%	98%	100%	100%	96%	100%	84%	88%	99%	95%	100%	72%	87%	98%	93%
	K=20, $n = 40$					K=20, $n = 80$					K=20, $n = 120$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	**0.58**	1.31	0.60	1.26	1.00	**0.67**	1.21	0.71	1.16	1.00	**0.69**	1.18	0.73	1.19
Share (%)	5%	45%	0%	**50%**	0%	6%	**53%**	5%	36%	0%	10%	**57%**	3%	30%	0%
Convergence(%)	100%	95%	97%	100%	88%	100%	93%	96%	100%	80%	100%	86%	96%	100%	70%

We report results from strong and weak VARMA(1,1) models simulated with different Kronecker indices, system sizes and fixed sample size set as $T = 200$. The upper part of the table refers to strong VARMA(1,1) models, while the lower part brings results from weak VARMA(1,1) models. The first column on each panel reports results from DGP III, while the second and third columns display results from models simulated with DGPs IV and V, respectively. Recall that DGP III sets the first Kronecker index to one and all the remaining indices to zero, $\mathbf{p} = (1, 0, 0, 0, ..., 0)'$; DGP IV sets the first two Kronecker indices to one and the remaining $K - 2$ indices to zero, $\mathbf{p} = (1, 1, 0, 0, ..., 0)'$; and DGP V sets the first three Kronecker indices to one and the remaining $K - 3$ indices to zero, $\mathbf{p} = (1, 1, 1, 0, ..., 0)'$. The true vectors of parameters in DGP III, IV and V contain the estimates obtained by fitting VARMA(1,1) models on the first rolling window of Dataset 1 in their respective system dimensions. Table S.2 displays the true values used to simulate DGPs III, IV and V. Weak innovations are generated using the procedure described in Romano and Thombs (1996). n accounts for the number of free parameters in the model. MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share % is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage of Hannan and Kavalieris (1984b); and KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012). The number of replications is set to 1000.

S. 6 Empirical Application

This section provides additional details on the BVAR framework and the Hannan-Kavalieris procedure implemented in Section 5. We also report the detailed results covering small ($K = 3$), medium, ($K = 10$ and $K = 20$), and large, ($K = 40$ and $K = 52$), sized systems. These results are reported in Tables S.10, S.11a, S.11b, S.12a, S.12a, S.13a and S.13b.

S. 6.1 BVAR

BVAR models have become an extremely popular approach to forecast key macroeconomic variables using large datasets, following the seminal articles of Doan et al. (1984), Litterman (1986) and Sims and Zha (1998). The BVAR framework builds on the idea of applying Bayesian shrinkage via the imposition of prior beliefs on the parameters of a K dimensional stable VAR(p) model,

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + ... + A_p Y_{t-p} + e_t. \tag{S.100}$$

The Minnesota prior shrinks the parameter estimates of (S.100) towards a random walk representation, so that the prior expectation and variance of the parameters are given by

$$\mathbb{E}\left[(A_k)_{ij}\right] = \begin{cases} \varpi, & j = i,\ k = 1 \\ 0, & \text{otherwise} \end{cases}, \qquad \text{Var}\left[(A_k)_{ij}\right] = \begin{cases} \frac{\varphi}{k^2} \\ \varphi\mu\frac{1}{k^2}\upsilon_i^2\upsilon_j^{-2}, \end{cases} \tag{S.101}$$

where φ is the hyperparameter that controls the overall tightness of the prior (amount of shrinkage), $\mu \in (0, 1)$ governs the importance of lags of other variables to the current values of the dependent variable, and ϖ is usually set in accordance with the persistence of the variables. Because the transformed variables we forecast are approximately covariance stationary processes, we follow the literature (Bańbura et al. (2010), Carriero et al. (2011) among others) and set $\varpi = 0$ and impose a white noise prior belief. The Minnesota prior has the drawback of setting the residual covariance matrix as a fixed diagonal matrix, such that $\text{Var}(e_t) = \Upsilon = \text{diag}(\upsilon_1^2, ..., \upsilon_K^2)$. Kadiyala and Karlsson (1997) propose a prior specification that retains the principle of the so-called Minnesota prior in (S.101) but relaxes the strong assumption of a fixed diagonal covariance matrix. The prior has a normal-inverted Wishart form

$$\text{vec}(B)|\Upsilon \sim \mathcal{N}\left(\text{vec}(B_0), \Upsilon \otimes \Omega_0\right), \qquad \Upsilon \sim iW\left(S_0, \alpha_0\right), \tag{S.102}$$

with moments given by

$$\mathbb{E}\left[(A_k)_{ij}\right] = 0, \qquad \text{Var}\left[(A_k)_{ij}\right] = \varphi \frac{1}{k^2} \upsilon_i^2 \upsilon_j^{-2}, \tag{S.103}$$

where $B = (A_1, ..., A_p)'$ and B_0, Ω_0, S_0 and α_0 are chosen such that the expectation of Υ is equal to the residual variance implied by the Minnesota prior, and the expectation and variance of B equal the moments implied by (S.101), with $\mu = 1$. We estimate the BVAR model with the normal-inverted Wishart prior by augmenting (S.100) with dummies variables (see Bańbura et al. (2010) and Kadiyala and Karlsson (1997) for more details). The hyperparameter φ plays a crucial role on the amount of shrinkage we impose on the parameter estimates and hence on the forecast performance of the BVAR models. When $\varphi = 0$, the prior is imposed exactly, while $\varphi = \infty$ yields the standard OLS estimates.

S. 6.2 Hannan-Kavalieris Procedure

We adopt the Hannan-Kavalieris procedure as discussed in Hannan and Kavalieris (1984a), Hannan and Kavalieris (1984b), Lütkepohl and Poskitt (1996) and Lütkepohl (2007) to specify the optimal Kronecker indices. The Hannan-Kavalieris procedure consists of minimizing an information criterion denoted by $C(\mathbf{p})$, given different alternative specifications of $\mathbf{p}$. Because estimation of VARMA models is usually demanding in medium- and high-dimensional systems, we follow (Lütkepohl, 2007, pg. 503) and adopt the HR estimator (first step of the IOLS estimator) when performing the Hannan-Kavalieris procedure. We can split the procedure into two steps. First, we start the procedure by exogenously defining the maximum value that the Kronecker indices may assume, which is denoted by $p_{\max}$. Following that, we estimate different VARMA models, assigning all elements of $\mathbf{p}$ to $p_{\max}, p_{\max} - 1, ..., 1$, successively. We choose $\mathbf{p}$ that minimizes the criterion $C(\mathbf{p})$, denoting this vector of Kronecker indices as $\mathbf{p}^{(1)} = \left(p_{(1)}, p_{(1)}, ..., p_{(1)}\right)'$, where $1 \leq p_{(1)} \leq p_{\max}$. The second step of the Hannan-Kavalieris procedure consists on choosing the Kronecker indices that will be used to estimate the model. This strategy requires $K\left(p_{(1)} + 1\right)$ evaluations, since we aim to define $p_k = \widehat{p}_k$, for all $k = 1, ..., K$ by minimizing the information criterion successively. The first evaluation requires the estimation of the model varying the K^{th} Kronecker index from $p_K = 0$ to $p_K = p_{(1)}$. We choose the $\widehat{p}_K$ associated with the lowest value of $C(\mathbf{p})$. At the end of this first evaluation, we have $\mathbf{p} = \left(p_1 = p_{(1)}, p_2 = p_{(1)}, ..., p_K = \widehat{p}_K\right)'$. We repeat the procedure for all of the remaining Kronecker indices. Therefore, the k^{th} evaluation results in the following vector of Kronecker indices: $\mathbf{p} = \left(p_1 = p_{(1)}, ..., p_{k-1} = p_{(1)}, p_k = \widehat{p}_k, \widehat{p}_{k+1}, ..., \widehat{p}_K\right)'$. Because we want to select the Kronecker indices which yield VARMA models that cannot be partitioned into smaller independent systems, we restrict $\widehat{p}_1 = 1$. This follows because if $\widehat{p}_1 = 0$ and $\widehat{p}_k = 0$ with $2 \leq k \leq \ell$, the necessary and sufficient restrictions imposed by the Echelon form transformation imply that first ℓ variables in the system are reduced to white noise processes and hence a potential high dimensional system can be partitioned into smaller independent systems. The information criterion is chosen to be the Schwarz criterion (SC), as it delivers consistent estimates of the Kronecker indices when U_t is a strong white noise process and the VARMA model is invertible and stable. The maximum value of the Kronecker indices is set to 1. Finally, it is important to note that the Hannan-Kavalieris procedure does not necessarily yield the minimum SC criterion over all possible combinations of Kronecker indices.

S. 6.3 Tables: Empirical Application

Table S.9: Dataset Specification

	K=3	$K = 10$				$K = 20$				$K = 40$				K=52
Dataset		1	2	3	4	1	2	3	4	1	2	3	4	
IPS10	×	×	×	×	×	×	×	×	×	×	×	×	×	×
FYFF	×	×	×	×	×	×	×	×	×	×	×	×	×	×
PUNEW	×	×	×	×	×	×	×	×	×	×	×	×	×	×
a0m052		×			×	×	×	×	×	×	×	×	×	×
A0M051							×				×	×	×	×
A0M224$_R$						×			×	×	×	×	×	×
A0M057		×				×				×	×		×	×
A0M059							×		×	×	×	×		×
PMP				×			×	×		×		×	×	×
A0m082					×					×	×	×	×	×
LHEL												×	×	×
LHELX						×			×	×	×		×	×
LHEM			×								×	×	×	×
LHUR		×				×		×	×	×			×	×
CES002				×				×		×		×	×	×
A0M048							×				×			×
PMI			×			×			×	×	×	×	×	×
PMNO					×					×		×		×
PMDEL				×				×				×	×	×
PMNV										×	×			×
FM1						×	×	×		×	×	×	×	×
FM2					×				×	×	×	×		×
FM3				×				×				×	×	×
FM2DQ		×				×		×	×	×	×	×	×	×
FMFBA			×							×	×	×		×
FMRRA								×				×	×	×
FMRNBA										×	×		×	×
FCLNQ					×		×		×	×	×		×	×
FCLBMC			×			×				×	×	×	×	×
CCINRV						×	×	×	×	×		×	×	×
A0M095						×	×			×	×	×		×
FSPCOM						×	×			×	×			×
FSPIN			×								×	×	×	×
FSDXP									×		×	×		×
FSPXE				×				×		×	×		×	×
CP90		×				×	×	×		×	×	×	×	×
FYGM3		×			×	×		×	×	×	×	×	×	×
FYGM6			×								×	×		×
FYGT1							×		×	×	×	×		×
FYGT5				×				×			×	×	×	×
FYGT10		×				×				×	×		×	×
FYAAAC							×		×			×	×	×
FYBAAC						×			×	×	×	×	×	×
EXRUS			×				×			×	×	×	×	×
EXRSW									×	×		×		×
EXRJAN							×		×	×	×		×	×
EXRUK							×			×	×	×	×	×
EXRCAN										×	×		×	×
PWFSA				×			×	×		×		×	×	×
PWFCSA					×					×	×	×	×	×
PWIMSA								×		×	×	×	×	×
PWCMSA						×		×		×	×	×	×	×

Table S.10: Forecast: Small and Large Sized Systems, $K = 3$ and $K = 52$

	$K=3$									$K=52$									
	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_1	BVAR$_{SC}$	BVAR$_{0.2}$	BVAR$_{opt}$	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	BVAR$_{SC}$	BVAR$_{0.2}$	BVAR$_{opt}$
Hor: 1																			
IPS10	0.87***	**0.86*** **	0.87***	0.87***	0.91**	0.92**	0.92**	0.92**	0.88**	0.99	0.98	0.96	6.36**	4.02***	0.90***	**0.89**	0.99	0.94	0.94
FYFF	0.89	**0.83***	0.85	0.92	1.17	1.05	1.11	1.09	0.91	1.97***	4.22***	4.08***	7.28**	20.45***	**0.89**	3.84***	4.13***	3.85***	3.97***
PUNEW	0.92	**0.91**	0.96	0.98	1.06	1.05	1.02	1.02	0.94	0.97	**0.94**	1.17	1.05	6.40***	1.02	1.07	1.22	1.16	1.20
Hor: 2																			
IPS10	0.91**	**0.90****	0.92***	0.92**	0.94	0.95	0.96	0.96	1.00	0.97	0.95	0.94	3.36*	4.59***	**0.89****	0.93	0.92	0.91	0.92
FYFF	0.91	0.90	**0.90**	0.91	1.31*	1.24*	1.17	1.15	0.98	1.19	1.27	1.22	3.08*	25.37***	1.00	1.84*	2.52***	2.28***	2.45***
PUNEW	**0.90***	0.91**	0.96	0.98	1.04	1.05	1.03	1.02	0.94*	0.92**	**0.91****	1.04	0.95	4.14***	1.04	1.06	1.05	1.03	1.03
Hor: 3																			
IPS10	0.92	**0.91***	0.93*	0.93*	0.97	0.98	0.98	0.98	1.01	0.98	0.99	0.99	1.77*	3.93***	**0.96**	1.04	1.00	1.01	1.00
FYFF	**0.79***	0.80*	0.85**	0.84	1.07	1.05	0.94	0.94	0.94*	0.93	0.88	0.86	1.59	24.10***	**0.79**	0.86	1.25	1.12	1.21
PUNEW	1.02	1.02	0.99	1.00	1.00	1.02	**0.97**	0.98	1.00	1.02	1.01	**0.90**	1.02	4.36***	1.00	1.02	0.97	0.96	0.96
Hor: 6																			
IPS10	0.94**	**0.93****	0.96**	0.95**	0.93***	1.00	0.99	0.99	1.00	0.99	0.99	1.00	1.11	3.53***	0.97	**0.91**	1.03	1.03	1.03
FYFF	**0.91**	0.91	0.95*	0.96	1.01	0.98	1.00	1.00	1.00	0.95	0.96	**0.93**	0.97	14.78***	1.27	1.09	1.00	0.95	0.98
PUNEW	**0.99**	0.99	0.99	0.99	0.99	1.07	0.99	0.99	0.99	1.00	1.00	1.00	0.98	2.70***	1.13	1.06	0.97	**0.97**	0.97
Hor: 9																			
IPS10	0.98*	0.98	0.99**	0.98**	0.99	**0.97**	1.00	1.00	1.00	0.99	1.00	1.00	1.00	2.12***	**0.93***	0.96	1.02	1.01	1.02
FYFF	0.97	0.97	0.99	0.99	0.99	**0.96**	1.01*	1.00	1.00	0.98	0.99	0.98	**0.93**	6.35***	0.94	1.18	0.96	0.96	0.96
PUNEW	1.00	1.00	1.00	1.00	1.00	**0.76****	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.92***	**0.77****	0.84	1.00	1.00	0.99
Hor: 12																			
IPS10	1.00	**1.00**	1.00	1.00	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.01	2.50***	**1.00**	1.01	1.02	1.02	1.01
FYFF	1.00	1.00	1.00	**0.99**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**0.96**	6.57***	1.06	1.40**	0.99	0.99	0.99
PUNEW	1.00	1.00	1.00	1.00	0.99	**0.82***	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.34	**0.88**	0.98	0.95	0.95	0.95
SC	-0.33	-0.38	-0.46	**-0.46**						**-61.35**	-60.67	-60.16	-43.99						
SC$_3$	-0.33	-0.38	-0.46	**-0.46**						**0.76**	1.81	2.92	17.26						
Convergence (%)	100%	100%	11%	37%						100%	100%	100%	0%						

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. All results are reported in terms of the RelMSFE. The RelMSFE is the ratio of the MSFE of a given model relative to the MSFE of the AR(1) model. The panel on the left-hand side reports results from small sized systems ($K = 3$), while the panel on the right-hand side displays results for large sized systems, $K = 52$. VARMA models denoted by $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ have the Kronecker indices $\mathbf{p} = (p_1, p_2, ..., p_K)'$ set as $p_i = 1$ for $i \leq k$ and $p_i = 0$ for all $i > k$ with $k = 1$, $k = 2$ and $k = 3$, respectively. $\mathbf{p}_{HK}$ is the VARMA model with Kronecker indices obtained using the Hannan-Kavalieris procedure. For $K = 3$, we implement the Hannan-Kavalieris on every rolling window, while for $K = 52$ we implement the Hannan-Kavalieris procedure on the first rolling window and carry the optimal Kronecker indices to the remaining rolling windows. All VARMA models are estimated using the IOLS algorithm. FM_1 is the factor model with one factor. FM_{IC_3} and FM_{SC} are factor models with number of factors defined using the IC_3 and the SC criteria, respectively. The maximum number of factors is equal to 4. We follow Stock and Watson (2002) and set the maximum lag length for the factor and the variable of interest equal to 3 and 6, respectively. VAR is the VAR(p^*) model with lag length p^* defined by the AIC criterion with the maximum lag length equal to 15 and 6 for $K = 3$ and $K = 52$, respectively. BVAR$_{SC}$, BVAR$_{0.2}$ and BVAR$_{opt}$ are Bayesian VAR models with the normal inverted Wishart prior which reproduces the principles of the Minnesota-type prior. The three sets of results differ in the way the hyperparameter φ is chosen. BVAR$_{SC}$ adopts φ which minimizes the SC criterion, BVAR$_{0.2}$ uses $\varphi = 0.2$, while BVAR$_{opt}$ is computed with the hyperparameter which minimizes, on every rolling window, the in-sample average of the one-step-ahead root mean squared forecast error of the K variables. SC$_K$ and SC$_3$ account for the SC criterion computed using the $(K \times K)$ covariance matrix of the residuals and its (3×3) upper block, respectively. The symbols *, **, and *** denote rejection, at the 10%, 5%, and 1% levels, of the null hypothesis of equal predictive accuracy according to the Diebold and Mariano (1995) test.

Table S.11a: Forecast: Medium Sized Systems, $K = 10$

	Dataset 1										Dataset 2									
	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	BVAR_{BIC}	$\text{BVAR}_{0.2}$	BVAR_{opt}	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	BVAR_{BIC}	$\text{BVAR}_{0.2}$	BVAR_{opt}
Hor: 1																				
IPS10	**0.88****	0.90**	0.91*	0.99	1.12	0.93***	1.00	0.96	0.96	0.90	0.84**	0.86**	0.88	**0.83****	0.96	0.89***	0.85**	0.85*	0.85*	0.86*
FYFF	1.09	1.08	1.12	1.12	1.40	**0.83**	0.89	1.07	1.06	0.87	0.96	1.61*	1.74***	1.31	1.62**	**0.84**	1.51*	1.57*	1.53*	1.46
PUNEW	0.92	0.93	1.04	0.96	1.12*	1.11**	1.10*	1.04	1.03	**0.91**	0.95	0.97	1.14	**0.94**	1.42**	0.97	1.19*	1.09	1.08	1.10
Hor: 2																				
IPS10	0.97	0.99	1.02	1.15**	1.15**	**0.92**	1.18***	1.09**	1.09**	1.09**	0.93	**0.91**	1.00	0.92	1.05	0.91*	1.11	0.95	0.95	0.97
FYFF	1.02	0.98	**0.96**	1.23**	1.51**	1.18*	1.31**	1.14	1.14	1.06	0.84	**0.76**	0.86	1.10	1.64**	0.92	1.46	1.08	1.05	1.00
PUNEW	**0.88*****	0.89***	0.93	0.92*	1.04	1.03	1.05	0.96	0.96	0.92**	0.88*	**0.87**	1.00	0.90*	1.13	0.99	1.02	0.96	0.95	0.95
Hor: 3																				
IPS10	**0.95**	0.96	1.00	1.16**	1.12*	0.95	1.19**	1.06	1.06	1.05*	0.94	0.93	1.03	1.01	0.97	0.98	**0.91**	0.98	0.98	0.98
FYFF	**0.86**	0.88	0.92	1.28***	1.34**	1.05	1.09	1.14	1.14	1.06	0.69*	**0.65****	0.73*	0.76	1.00	0.79	0.85	0.71*	0.71*	0.71*
PUNEW	1.00	1.00	**0.91**	1.02	0.99	0.98	1.00	0.97	0.97	1.01	1.01	0.99	**0.95**	1.01	1.04	1.04	1.05	0.97	0.97	0.96
Hor: 6																				
IPS10	**0.93*****	0.94***	0.96	1.09	1.12*	1.03**	1.04	1.02	1.02	1.00	**0.93**	0.94	1.00	0.97	1.00	0.99	0.93	0.99	0.99	1.01
FYFF	**0.93**	0.95	1.02	1.22***	1.19***	1.29**	1.37*	1.12***	1.12***	1.01**	**0.80**	0.82	1.04	0.85	0.95	1.00	0.98	0.82	0.82	0.84
PUNEW	0.99	0.99	**0.98**	1.00	1.00	1.15	1.12	1.00	1.00	1.00	0.98	0.98	**0.95**	0.98*	0.95	1.03	1.13	0.96	0.96	0.95
Hor: 9																				
IPS10	0.98	0.98	1.00	1.02	1.05	**0.96**	1.03	1.01	1.01	1.00	0.97	0.98	1.01	0.99	0.99	**0.93**	0.96	1.02	1.02	1.02
FYFF	0.96	0.98	1.01	1.09**	1.07**	**0.91**	0.98	1.04*	1.04*	1.00	**0.92**	0.94	1.08	0.96*	0.93	0.98	1.23	0.93	0.93	0.94
PUNEW	1.00	1.00	1.00	1.00	1.00	**0.76****	0.79*	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.97	**0.78****	0.83	0.99	0.99	0.99
Hor: 12																				
IPS10	1.00	1.00	1.00	**0.99**	1.02	1.00	1.04	1.00	1.00	1.00	1.01	1.00	1.00	1.01	1.01	1.00	1.04	1.06*	1.06*	1.06**
FYFF	1.00	1.00	1.00	1.00	1.05	1.12**	1.16***	1.00	1.00	**1.00**	0.99	0.99	1.02	1.01	**0.98**	1.00	1.12	1.02	1.02	1.01
PUNEW	1.00	1.00	0.98***	0.99*	0.99	**0.85***	0.98	1.00	1.00	1.00	0.99	1.00	0.98**	1.00	0.95	**0.82***	0.91	1.00	1.00	1.01
SC	-4.89	-5.00	-5.05	**-5.52**							-3.12	-2.99	-3.20	**-3.89**						
SC_3	**-0.22**	-0.17	-0.09	0.04							**-0.29**	-0.27	-0.19	-0.07						
Convergence (%)	100%	100%	100%	100%							100%	100%	100%	97%						

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. The RelMSFE is the ratio of the MSFE of a given model against the MSFE of the AR(1) model. VARMA models denoted by $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ have Kronecker indices equal to $(1,0,0,0,....,0)'$, $(1,1,0,0,....,0)'$, $(1,1,1,0,....,0)'$, respectively. $\mathbf{p}_{HK}$ is a VARMA model with Kronecker indices obtained using the Hannan-Kavalieris procedure on every rolling window. All VARMA models are estimated using the IOLS algorithm. FM_{IC_3} and FM_{SC} are factor models with number of factors defined using the IC_3 and the SC criteria, where the maximum number of factors is equal to 4. We follow Stock and Watson (2002) and set the maximum lag length for the factor and the variable of interest equal to 3 and 6, respectively. VAR is the VAR(p^*) model with lag length p^* defined by the AIC criterion, with the maximum lag length equal to 15. BVAR_{SC}, $\text{BVAR}_{0.2}$ and BVAR_{opt} are Bayesian VAR models with the normal inverted Wishart prior which reproduces the principles of the Minnesota-type prior. The three sets of results differ in the way the hyperparameter φ is chosen. BVAR_{SC} adopts φ which minimizes the Schwarz criterion, $\text{BVAR}_{0.2}$ uses $\varphi = 0.2$, while BVAR_{opt} is computed with the hyperparameter which minimizes, on every rolling window, the in-sample average of the one-step-ahead root mean squared forecast error of the K variables. SC_K and SC_3 account for the Schwarz criteria computed using the $(K \times K)$ covariance matrix of the residuals and its (3×3) upper block, respectively. The symbols *, **, and *** denote rejection, at the 10%, 5%, and 1% levels, of the null hypothesis of equal predictive accuracy according to the Diebold and Mariano (1995) test.

Table S.11b: Forecast: Medium Sized Systems, $K = 10$

	Dataset 3										Dataset 4									
	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	$BVAR_{BIC}$	$BVAR_{0.2}$	$BVAR_{opt}$	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	$BVAR_{BIC}$	$BVAR_{0.2}$	$BVAR_{opt}$
Hor: 1																				
IPS10	0.79**	0.79**	0.80**	0.82**	0.83*	0.97	0.81***	0.78***	0.78***	**0.77*** **	0.76***	0.78**	0.82*	0.80**	0.87	0.90***	0.91	**0.75*** **	0.75***	0.76***
FYFF	**0.89**	1.98**	1.93**	1.95**	2.88***	0.93	1.33	1.97**	1.93**	1.66	1.17	1.48*	1.50*	1.54**	1.58**	**0.79**	1.10	1.27	1.26	1.10
PUNEW	0.98	1.01	1.02	1.02	1.16	0.97	**0.86***	0.91	0.91	0.90	0.96	0.97	0.96	0.94	1.08	1.06	0.93	**0.86****	0.86**	0.87*
Hor: 2																				
IPS10	0.89	**0.88**	0.94	0.93	0.91	0.96	1.06	0.89	0.89	0.91	0.91	0.91	1.00	0.92	1.00	0.92*	0.95	0.91	0.91	**0.90**
FYFF	0.75	**0.67***	0.72*	0.79	1.68*	1.06	0.88	1.03	1.01	0.89	0.82	**0.79**	1.02	0.94	1.21	1.06	1.13	1.05	1.04	0.93
PUNEW	**0.91***	0.92*	1.12	1.06	1.14	1.01	1.02	0.97	0.97	0.96	**0.91***	0.91*	1.07	0.93	1.16	1.07	1.11	1.01	1.01	1.00
Hor: 3																				
IPS10	0.95	**0.94**	1.00	1.01	0.98	0.96	1.09	0.95	0.94	0.96	0.94	**0.93**	1.04	0.97	1.04	0.96	1.05	0.98	0.98	0.97
FYFF	**0.66****	0.66**	0.68**	0.75	1.23	0.89	0.86	0.72	0.71	0.73*	0.70*	**0.69****	0.90	0.83	0.86	0.85	0.90	0.87	0.87	0.83
PUNEW	1.00	0.99	**0.94**	1.00	1.08	1.02	1.04	0.97	0.97	0.97	1.01	1.01	**0.98**	1.02	0.99	0.99	1.00	0.99	0.99	0.99
Hor: 6																				
IPS10	0.95	0.96	0.99	0.93	0.93	1.02	1.02	**0.93**	0.93	0.96	0.95	0.95	1.01	0.96	0.98	1.01*	0.96	**0.95**	0.95	0.96
FYFF	0.82	0.86	0.92*	0.83	0.96	0.96	1.34**	0.79	**0.79**	0.83	**0.83**	0.85	1.06	0.85	0.87	0.92	0.89	0.87	0.87	0.89
PUNEW	0.98	0.99	0.99	0.97	**0.91**	1.01	1.06	0.96	0.96	0.95	0.99	0.99	0.99	0.97	0.98	1.03	1.13	0.97	0.97	**0.97**
Hor: 9																				
IPS10	0.98	0.98	1.01	0.95	**0.94**	0.97	0.99	0.96	0.96	0.98	0.99	0.99	1.01	0.99	0.99	**0.98**	1.00	0.99	0.99	0.99
FYFF	0.92	0.95	1.00	0.89	**0.85**	0.97	1.00	0.88	0.89	0.91	0.93	0.94	1.06	**0.90**	0.92	0.93	0.92	0.91	0.91	0.92
PUNEW	0.99	1.00	1.00	1.03	0.98	**0.77****	0.82*	1.01	1.01	1.01	0.99	0.99	0.98	0.98	0.97	0.78**	**0.77****	0.98	0.98	0.98
Hor: 12																				
IPS10	1.01	1.01	**0.99**	1.00	1.00	1.02	1.07	1.01	1.01	1.02*	1.01	1.00	1.00	1.03	1.02	1.01	1.01	1.02	1.02	1.02
FYFF	0.99	0.99	1.01*	**0.98**	0.99	0.99	1.17**	0.99	0.99	1.01	0.99	0.99	1.01	0.99	**0.98**	1.00	1.14	0.98	0.98	0.99
PUNEW	1.00	1.00	0.98**	1.01	1.02	**0.84**	0.93	1.03	1.03	1.04	1.00	1.00	0.98***	0.99	0.97	**0.84**	0.87	0.99	0.99	0.99
SC	-3.95	-3.92	-4.40	**-5.86**							-5.65	-5.60	-5.69	**-6.15**						
SC_3	**-0.28**	-0.24	-0.21	0.36							**-0.32**	-0.28	-0.24	0.14						
Convergence (%)	100%	100%	100%	100%							100%	100%	100%	100%						

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. The RelMSFE is the ratio of the MSFE of a given model against the MSFE of the AR(1) model. VARMA models denoted by $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ have Kronecker indices equal to $(1,0,0,0,....,0)'$, $(1,1,0,0,....,0)'$, $(1,1,1,0,....,0)'$, respectively. $\mathbf{p}_{HK}$ is a VARMA model with Kronecker indices obtained using the Hannan-Kavalieris procedure on every rolling window. All VARMA models are estimated using the IOLS algorithm. FM_{IC_3} and FM_{SC} are factor models with number of factors defined using the IC_3 and the SC criteria, where the maximum number of factors is equal to 4. We follow Stock and Watson (2002) and set the maximum lag length for the factor and the variable of interest equal to 3 and 6, respectively. VAR is the VAR(p^*) model with lag length p^* defined by the AIC criterion, with the maximum lag length equal to 15. $BVAR_{SC}$, $BVAR_{0.2}$ and $BVAR_{opt}$ are Bayesian VAR models with the normal inverted Wishart prior which reproduces the principles of the Minnesota-type prior. The three sets of results differ in the way the hyperparameter φ is chosen. $BVAR_{SC}$ adopts φ which minimizes the Schwarz criterion, $BVAR_{0.2}$ uses $\varphi = 0.2$, while $BVAR_{opt}$ is computed with the hyperparameter which minimizes, on every rolling window, the in-sample average of the one-step-ahead root mean squared forecast error of the K variables. SC_K and SC_3 account for the Schwarz criteria computed using the $(K \times K)$ covariance matrix of the residuals and its (3×3) upper block, respectively. The symbols *, **, and *** denote rejection, at the 10%, 5%, and 1% levels, of the null hypothesis of equal predictive accuracy according to the Diebold and Mariano (1995) test.

Table S.12a: Forecast: Medium Sized Systems, $K = 20$

	Dataset 1										Dataset 2									
	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	BVAR_{BIC}	$\text{BVAR}_{0.2}$	BVAR_{opt}	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	BVAR_{BIC}	$\text{BVAR}_{0.2}$	BVAR_{opt}
Hor: 1																				
IPS10	0.92	0.95	0.94	**0.90**	0.96	0.91***	0.91	0.96	0.96	0.96	0.87	0.91	0.91	0.87	0.86	0.92***	0.91*	0.85	0.85	**0.82****
FYFF	1.31	4.30***	4.30***	1.48***	3.93***	1.03	1.52**	3.93***	3.87***	3.88***	1.66**	3.47***	3.71***	1.67**	4.23***	1.03	1.14	3.71***	3.66***	2.96***
PUNEW	**0.95**	0.99	1.07	0.95	1.02	1.02	1.08	1.02	1.01	1.00	0.99	1.01	1.04	0.99	1.07	1.03	1.08	0.95	0.94	**0.89***
Hor: 2																				
IPS10	0.95	0.95	1.02	1.04	0.97	**0.91***	1.02	0.97	0.97	0.97	1.02	1.01	1.06	0.96	0.98	**0.90***	0.96	1.01	1.00	0.97
FYFF	0.98	0.92	**0.90**	1.27*	1.86***	1.00	1.32*	1.86***	1.83***	1.84***	0.92	0.93	0.98	**0.69***	2.27***	1.04	1.19	1.76**	1.75**	1.47
PUNEW	**0.90***	0.91*	1.09	0.92*	1.06	1.02	1.03	1.06	1.06	1.06	0.91*	0.92*	1.09	**0.90****	1.17**	1.06	1.06	1.05	1.04	1.01
Hor: 3																				
IPS10	0.98	0.99	1.07	1.10	1.00	**0.95**	1.00	1.00	1.00	1.00	0.99	0.97	1.03	0.95	0.95	0.93	**0.90**	0.96	0.96	0.94
FYFF	0.82	0.78	0.81	1.16	0.97	0.81**	0.93	0.97	0.96	0.94	0.82	0.83	0.86	**0.62***	1.11	1.15***	1.10	0.99	0.98	0.93
PUNEW	0.99	1.00	0.98	1.03	0.98	0.99	1.03	0.98	0.98	**0.98**	1.00	0.99	0.96	0.99	0.96	1.03	0.99	0.96	**0.96**	0.98
Hor: 6																				
IPS10	**0.94**	0.94	0.98	1.07	1.01	1.01	0.99	1.01	1.01	1.01	0.97	0.97	1.00	0.96	0.97	0.98	1.01	**0.94**	0.94	0.97
FYFF	**0.88**	0.88	0.98	1.24***	0.98	1.14**	1.41	0.98	0.98	0.97	0.90	0.90	0.96	**0.73**	0.88	1.18**	1.20*	0.87	0.86	0.89
PUNEW	0.98	0.98	0.98	0.99	0.96	1.04	1.14	**0.96**	0.96	0.96	0.99	0.99	0.99	0.96	**0.94**	1.09	1.08	0.95	0.95	0.95
Hor: 9																				
IPS10	0.98	0.98	1.00	1.06	1.02	**0.95***	0.98	1.02	1.02	1.01	0.99	0.99*	1.00	1.01	0.98	**0.97**	1.03	0.98	0.98	1.00
FYFF	**0.91**	0.92	1.01	1.20**	0.99	0.94	0.99	0.99	0.99	0.99	0.97	0.97	1.01	**0.89**	0.92	1.05	1.10	0.92	0.92	0.95
PUNEW	1.00	1.00	1.01	1.00	1.00	**0.74****	0.84*	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	**0.74****	0.80**	0.99	0.99	0.99
Hor: 12																				
IPS10	1.02	1.01	1.00	1.01	1.03*	**0.99**	1.03	1.03*	1.03*	1.03*	1.00	1.00	1.00	1.08*	1.01	**0.99**	0.99	1.03	1.03*	1.02
FYFF	0.99	**0.99**	1.01	1.07	1.02	1.10*	1.19*	1.02	1.02	1.01	0.99	1.00	1.00	1.08	**0.99**	1.08	1.06	0.99	0.99	1.00
PUNEW	1.00	1.00	0.98**	0.99*	0.99	**0.81****	0.91	0.99	0.99	0.99	1.00	1.00	0.99***	1.00	0.97	**0.81***	0.86	0.99	0.99	1.00
SC	-10.47	-10.43	-10.59	**-11.13**							-11.32	-11.09	-10.94	**-11.44**						
SC_3	**-0.06**	0.21	0.49	1.43							**-0.06**	0.19	0.47	0.39						
Convergence (%)	100%	100%	100%	0%							100%	58%	95%	100%						

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. The RelMSFE is the ratio of the MSFE of a given model against the MSFE of the AR(1) model. VARMA models denoted by $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ have Kronecker indices equal to $(1,0,0,0,....,0)'$, $(1,1,0,0,....,0)'$, $(1,1,1,0,....,0)'$, respectively. $\mathbf{p}_{HK}$ is a VARMA model with Kronecker indices obtained using the Hannan-Kavalieris procedure on the first rolling window. All VARMA models are estimated using the IOLS algorithm. FM_{IC_3} and FM_{SC} are factor models with number of factors defined using the IC_3 and the SC criteria, where the maximum number of factors is equal to 4. We follow Stock and Watson (2002) and set the maximum lag length for the factor and the variable of interest equal to 3 and 6, respectively. VAR is the VAR(p^*) model with lag length p^* defined by the AIC criterion, with the maximum lag length equal to 8. BVAR_{SC}, $\text{BVAR}_{0.2}$ and BVAR_{opt} are Bayesian VAR models with the normal inverted Wishart prior which reproduces the principles of the Minnesota-type prior. The three sets of results differ in the way the hyperparameter φ is chosen. BVAR_{SC} adopts φ which minimizes the Schwarz criterion, $\text{BVAR}_{0.2}$ uses $\varphi = 0.2$, while BVAR_{opt} is computed with the hyperparameter which minimizes, on every rolling window, the in-sample average of the one-step-ahead root mean squared forecast error of the K variables. SC_K and SC_3 account for the Schwarz criteria computed using the $(K \times K)$ covariance matrix of the residuals and its (3×3) upper block, respectively. The symbols *, **, and *** denote rejection, at the 10%, 5%, and 1% levels, of the null hypothesis of equal predictive accuracy according to the Diebold and Mariano (1995) test.

Table S.12b: Forecast: Medium Sized Systems, $K = 20$

	Dataset 3										**Dataset 4**									
	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	$BVAR_{BIC}$	$BVAR_{0.2}$	$BVAR_{opt}$	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	$BVAR_{BIC}$	$BVAR_{0.2}$	$BVAR_{opt}$
Hor: 1																				
IPS10	0.85*	0.86	0.86	0.91	1.01	0.91***	0.86*	0.85*	0.84*	**0.82****	0.92	0.94	0.95	0.91	0.96	**0.87*****	0.99	0.96	0.94	0.93
FYFF	1.62***	2.89***	2.89***	3.16***	3.72***	1.09	4.00*	3.16***	3.02***	1.99***	1.05	1.59*	1.56*	1.07	1.51*	1.32	1.45*	1.51*	1.49	1.17
PUNEW	0.95	0.96	1.04	0.96	1.09	0.96	1.14	0.99	0.98	**0.91**	0.94	0.97	0.96	**0.93**	0.94	0.99	0.98	0.94	0.94	0.95
Hor: 2																				
IPS10	1.01	1.02	1.04	1.08	1.15	**0.94**	1.22	1.01	1.01	1.01	0.91	**0.90***	0.91	0.91	0.96	0.93	1.05	0.96	0.96	0.98
FYFF	1.13	0.98	0.94	**0.92**	2.61***	1.08	1.98	1.98**	1.92**	1.57**	0.83	0.72	**0.72***	0.92	1.14	0.94	1.22	1.14	1.12	1.01
PUNEW	0.92*	0.92*	1.10	**0.92****	1.18*	1.01	1.17	1.04	1.03	0.97	**0.90****	0.90**	1.05	0.91**	1.03	1.03	1.03	1.03	1.03	1.00
Hor: 3																				
IPS10	1.03	1.02	1.05	1.14	1.16	**0.97**	1.24	1.02	1.02	1.00	0.97	**0.95**	0.99	0.98	1.01	0.97	1.07	1.01	1.01	0.99
FYFF	0.85	0.82	**0.79**	0.90	1.66***	0.82**	0.97	0.98	0.97	0.95	0.75	**0.71****	0.73**	0.83	0.90	1.05	1.06	0.90	0.89	0.89
PUNEW	1.00	0.99	**0.94**	1.00	0.96	0.98	1.00	0.97	0.97	0.97	1.01	1.00	0.99	1.02	1.00	1.01	**0.96**	1.00	1.00	0.99
Hor: 6																				
IPS10	**0.97**	0.98	0.99	1.06	1.09	0.99	1.03	0.98	0.98	0.99	0.96	0.96	0.98	**0.95**	1.02	1.00	1.00	1.02	1.01	1.00
FYFF	**0.88**	0.89	0.92	1.03	1.06	1.06	1.28	0.89	0.88	0.88	**0.90**	0.91	0.96	0.94	0.95	1.27**	1.17	0.95	0.94	0.90
PUNEW	0.98	0.99	0.99	0.99	**0.91**	1.00	1.05	0.96	0.96	0.95	0.99	0.99	0.98	0.99	0.93	1.05	1.07	**0.93**	0.93	0.95
Hor: 9																				
IPS10	0.98	0.98	1.00	1.00	1.02	**0.94**	1.04	0.99	0.99	1.00	0.99	0.99	1.00	0.99	1.02	**0.95***	1.00	1.02	1.02	1.01
FYFF	**0.93**	0.94	0.98	1.01	1.00	0.98	1.14	0.94	0.94	0.96	0.96	0.97	1.01	**0.95**	0.99	1.00	1.16	0.99	0.99	0.95
PUNEW	1.00	1.00	1.00	1.00	0.99	**0.76****	0.81*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**0.73****	0.79*	1.00	1.00	0.99
Hor: 12																				
IPS10	1.01	1.01	1.00	**0.99**	1.02	1.03	1.08	1.01	1.01	1.00	1.00	1.00	1.00	1.01	1.04**	0.99	**0.98**	1.04**	1.04**	1.03*
FYFF	**0.99**	0.99	1.00	1.00	1.01	1.04	1.27***	1.00	1.00	1.01	1.00	1.00	1.00	**0.99**	1.01	1.13**	1.15	1.01	1.01	1.00
PUNEW	1.00	1.00	0.99**	0.99***	0.96	**0.82***	0.93	0.99	0.99	1.01	1.00	1.00	1.00*	1.00	1.00	**0.81****	0.92	1.00	1.00	1.00
SC	-10.30	-10.23	-10.36	**-12.13**							-11.34	-11.23	-11.44	**-12.54**						
SC_3	**-0.05**	0.21	0.46	2.36							**-0.06**	0.18	0.48	0.91						
Convergence (%)	100%	100%	100%	95%							100%	100%	100%	100%						

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. The RelMSFE is the ratio of the MSFE of a given model against the MSFE of the AR(1) model. VARMA models denoted by $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ have Kronecker indices equal to $(1,0,0,0,....,0)'$, $(1,1,0,0,....,0)'$, $(1,1,1,0,....,0)'$, respectively. $\mathbf{p}_{HK}$ is a VARMA model with Kronecker indices obtained using the Hannan-Kavalieris procedure on the first rolling window. All VARMA models are estimated using the IOLS algorithm. FM_{IC_3} and FM_{SC} are factor models with number of factors defined using the IC_3 and the SC criteria, where the maximum number of factors is equal to 4. We follow Stock and Watson (2002) and set the maximum lag length for the factor and the variable of interest equal to 3 and 6, respectively. VAR is the VAR(p^*) model with lag length p^* defined by the AIC criterion, with the maximum lag length equal to 8. $BVAR_{SC}$, $BVAR_{0.2}$ and $BVAR_{opt}$ are Bayesian VAR models with the normal inverted Wishart prior which reproduces the principles of the Minnesota-type prior. The three sets of results differ in the way the hyperparameter φ is chosen. $BVAR_{SC}$ adopts φ which minimizes the Schwarz criterion, $BVAR_{0.2}$ uses $\varphi = 0.2$, while $BVAR_{opt}$ is computed with the hyperparameter which minimizes, on every rolling window, the in-sample average of the one-step-ahead root mean squared forecast error of the K variables. SC_K and SC_3 account for the Schwarz criteria computed using the $(K \times K)$ covariance matrix of the residuals and its (3×3) upper block, respectively. The symbols *, **, and *** denote rejection, at the 10%, 5%, and 1% levels, of the null hypothesis of equal predictive accuracy according to the Diebold and Mariano (1995) test.

Table S.13a: Forecast: Large Sized Systems, $K = 40$

	Dataset 1										Dataset 2									
	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	BVAR_{BIC}	$\text{BVAR}_{0.2}$	BVAR_{opt}	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	BVAR_{BIC}	$\text{BVAR}_{0.2}$	BVAR_{opt}
Hor: 1																				
IPS10	0.86	0.91	0.88	3.79	2.91***	0.89***	0.87**	0.85	0.85*	**0.84***	0.97	0.98	0.94	0.98	3.58***	**0.85*****	0.87*	0.96	0.91	0.93
FYFF	1.53**	4.11***	3.99***	4.37	19.27***	**0.94**	2.04*	3.80***	3.70***	3.65***	1.82***	3.74***	3.67***	3.83***	22.30***	1.01	2.65***	3.56***	3.29***	3.31***
PUNEW	**0.97**	0.97	1.14	0.97	3.40***	0.98	0.98	1.12	1.11	1.12	0.99	**0.96**	1.27**	0.97	3.33***	0.98	0.98	1.27**	1.21	1.22*
Hor: 2																				
IPS10	0.94	0.92	0.91	2.38	2.60***	0.89**	0.97	0.86*	**0.86***	0.86*	0.99	0.98	1.00	0.95	3.34***	0.90	**0.88**	0.92	0.91	0.91
FYFF	1.08	1.06	1.01	1.97	23.41***	**0.85**	1.40	2.01***	1.93***	2.00***	1.13	1.10	1.06	1.16	19.48***	1.02	1.64**	1.84**	1.67**	1.87**
PUNEW	0.93*	0.93*	1.07	**0.91****	3.64***	1.02	1.14	1.07	1.06	1.06	0.94	**0.93***	1.07	0.93	3.45***	1.02	1.10	1.04	1.03	1.01
Hor: 3																				
IPS10	1.01	1.00	1.01	1.97	2.82***	**0.96**	0.96	0.96	0.96	0.96	1.00	1.01	1.04	0.98	3.28***	**0.94**	0.96	0.96	0.97	0.99
FYFF	0.87	0.82**	0.79**	1.58	16.48***	**0.70****	1.03	1.08	1.03	1.08	0.91	0.83*	**0.81***	0.83	12.59***	1.13***	0.84	1.05	0.96	1.08
PUNEW	1.02	1.01	**0.93**	1.03	4.02**	0.99	0.98	0.99	0.99	0.99	1.01	1.01	**0.94**	1.01	3.82***	1.01	1.02	1.02	1.01	1.01
Hor: 6																				
IPS10	0.98	0.98	1.00	1.54	2.78***	1.01	**0.91**	0.97	0.97	0.98	0.97	0.98	1.00	0.98	2.80***	0.96	**0.93**	0.98	0.98	0.99
FYFF	0.95	0.96	0.95	1.49	10.85***	1.00	1.00	0.93	**0.92**	0.93	**0.93**	0.94	0.96	0.95	10.26***	1.27***	1.14	0.98	0.95	0.99
PUNEW	1.00	1.00	1.02	1.00	2.23***	0.98	1.13	0.98	**0.98**	0.98	0.99	1.00	1.00	1.00	2.08***	1.06	1.12	**0.94**	0.95	0.95
Hor: 9																				
IPS10	1.00	1.00	1.00	1.23	1.96*	0.97	**0.94**	0.99	0.99	0.99	0.99*	0.99	1.00	0.99	2.41***	**0.92****	0.94	0.99	0.99	1.00
FYFF	0.97	0.99	0.98	1.21	5.26***	0.95	1.26	0.93	0.93	**0.93**	**0.97**	0.98	0.99	0.99	7.58***	0.98	0.98	0.98	0.98	0.99
PUNEW	1.00	1.00	0.99	0.99*	1.61***	**0.76****	0.81*	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.68***	**0.74****	0.82	0.99	0.99	0.98
Hor: 12																				
IPS10	1.00	1.00	1.00	1.05	1.85***	1.02	**0.98**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.02***	**0.99**	1.02	1.00	1.00	1.00
FYFF	1.00	1.00	1.00	1.02	4.74***	1.04	1.25*	**0.98**	0.98	0.98	**1.00**	1.00	1.00	1.00	4.21***	1.07	1.17*	1.03	1.02	1.02
PUNEW	1.00	1.00	1.00	0.99**	1.19	**0.83**	0.94	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.49**	**0.80****	1.05	0.97	0.97	0.97
SC	**-35.96**	-35.42	-35.23	-33.62							**-41.65**	-41.16	-40.87	-40.29						
SC_3	**0.47**	1.21	2.00	7.42							**0.50**	1.23	2.06	2.35						
Convergence (%)	100%	100%	100%	74%							100%	100%	100%	100%						

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. The RelMSFE is the ratio of the MSFE of a given model against the MSFE of the AR(1) model. VARMA models denoted by $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ have Kronecker indices equal to $(1,0,0,0,\ldots,0)'$, $(1,1,0,0,\ldots,0)'$, $(1,1,1,0,\ldots,0)'$, respectively. $\mathbf{p}_{HK}$ is a VARMA model with Kronecker indices obtained using the Hannan-Kavalieris procedure on the first rolling window. All VARMA models are estimated using the IOLS algorithm. FM_{IC_3} and FM_{SC} are factor models with number of factors defined using the IC_3 and the SC criteria, where the maximum number of factors is equal to 4. We follow Stock and Watson (2002) and set the maximum lag length for the factor and the variable of interest equal to 3 and 6, respectively. VAR is the VAR(p^*) model with lag length p^* defined by the AIC criterion, with the maximum lag length equal to 8. BVAR_{SC}, $\text{BVAR}_{0.2}$ and BVAR_{opt} are Bayesian VAR models with the normal inverted Wishart prior which reproduces the principles of the Minnesota-type prior. The three sets of results differ in the way the hyperparameter φ is chosen. BVAR_{SC} adopts φ which minimizes the Schwarz criterion, $\text{BVAR}_{0.2}$ uses $\varphi = 0.2$, while BVAR_{opt} is computed with the hyperparameter which minimizes, on every rolling window, the in-sample average of the one-step-ahead root mean squared forecast error of the K variables. SC_K and SC_3 account for the Schwarz criteria computed using the $(K \times K)$ covariance matrix of the residuals and its (3×3) upper block, respectively. The symbols *, **, and *** denote rejection, at the 10%, 5%, and 1% levels, of the null hypothesis of equal predictive accuracy according to the Diebold and Mariano (1995) test.

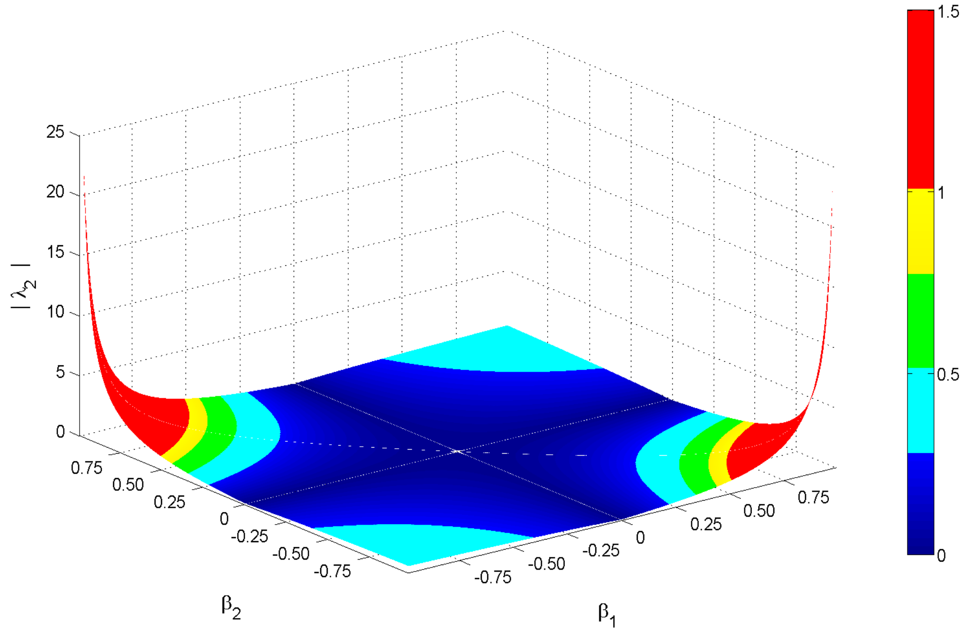
Table S.13b: Forecast: Large Sized Systems, $K = 40$

	Dataset 3										Dataset 4									
	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	$BVAR_{BIC}$	$BVAR_{0.2}$	$BVAR_{opt}$	$\mathbf{p}_{100}$	$\mathbf{p}_{110}$	$\mathbf{p}_{111}$	$\mathbf{p}_{HK}$	VAR	FM_{SC}	FM_{IC_3}	$BVAR_{BIC}$	$BVAR_{0.2}$	$BVAR_{opt}$
Hor: 1																				
IPS10	0.90	0.90	0.89	3.63***	2.57***	0.87***	**0.86****	0.90	0.89	0.90	1.00	1.00	0.98	2.30***	3.42***	**0.90*****	0.93	1.01	1.00	0.99
FYFF	1.84***	3.32***	3.23***	26.84***	21.10***	**1.00**	3.46**	3.23***	3.05***	3.13***	1.70***	4.93***	4.91***	2.17***	19.42***	**0.96**	3.51*	5.13***	4.88***	4.92***
PUNEW	**0.95**	0.96	1.02	1.02	3.76***	1.00	1.00	1.05	1.04	1.05	**0.95**	0.98	1.06	0.98	3.41***	1.01	1.12	1.02	1.02	1.02
Hor: 2																				
IPS10	0.94	0.92	0.92	2.06	3.33***	**0.90****	0.98	0.99	0.98	0.99	0.98	0.92	0.92	1.00	2.76***	**0.89****	1.02	0.95	0.95	0.93
FYFF	1.11	1.22	1.20	9.04***	23.33***	**0.90**	1.89	2.31***	2.10***	2.30***	1.21	1.09	1.03	**0.72***	27.19***	0.93	2.80	2.71***	2.54***	2.67***
PUNEW	0.92**	**0.92****	1.04	1.05	3.58***	1.03	1.14	1.00	1.00	1.00	**0.92****	0.92**	1.09	0.93**	3.56***	1.02	1.15	1.05	1.04	1.05
Hor: 3																				
IPS10	**0.96**	0.98	0.99	1.96*	2.46***	0.97	1.11	1.01	1.01	1.01	1.05	1.01	1.03	**0.92**	2.88***	0.97	1.24	1.03	1.03	1.01
FYFF	0.90	0.87	0.86	15.09*	18.02***	**0.82****	0.91	1.15	1.08	1.11	0.94	0.84	0.81	**0.71****	21.13**	0.76**	1.08	1.08	1.04	1.02
PUNEW	1.02*	1.01	**0.91**	1.12**	3.36***	1.00	1.03	0.96	0.95	0.96	1.01	1.00	**0.97**	1.01	3.51***	1.00	1.00	0.98	0.98	0.98
Hor: 6																				
IPS10	0.98	0.99	1.00	1.00	2.74***	1.00	**0.96**	1.05	1.04	1.05	0.99	0.99	0.99	0.98	2.33***	1.01	**0.96**	1.01	1.01	1.00
FYFF	0.95	0.95	**0.93**	1.51	9.84***	1.20**	1.05	0.98	0.97	0.99	0.94	0.93	0.91	**0.84**	11.50**	1.08*	1.08	0.92	0.91	0.89
PUNEW	1.00	1.00	1.01	1.00	1.75***	**0.97**	1.10	0.99	0.99	0.99	0.99	1.00	0.99	0.99	2.80**	0.99	1.11	0.97	**0.96**	0.97
Hor: 9																				
IPS10	0.99	1.00	1.00	**0.96**	2.27***	0.97	1.00	1.01	1.01	1.03	0.99	0.99	1.00	0.97	1.72**	**0.97**	1.02	1.01	1.01	1.01
FYFF	0.98	0.99	0.98	1.00	7.60***	**0.94**	1.21	0.97	0.96	0.96	0.98	0.98	0.97	0.94	7.20***	**0.93**	1.12	0.96	0.95	0.95
PUNEW	1.00	1.00	1.00	0.99	1.52***	**0.74****	0.81*	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.66*	**0.75****	0.84	1.00	1.00	0.99
Hor: 12																				
IPS10	1.00	1.00	1.00	1.00	1.97**	**0.99**	1.02	1.01	1.01	1.01	1.00	1.00	1.00	**1.00**	2.43**	1.01	1.06	1.01	1.01	1.01
FYFF	1.00	1.00	1.00	1.00	5.90***	1.09**	1.32***	**0.99**	0.99	0.99	1.00	1.00	1.00	**0.97**	7.36*	1.07*	1.37***	1.01	1.00	1.00
PUNEW	1.00	1.00	1.00	1.00	1.39**	**0.82***	0.95	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.42	**0.83**	1.11	0.98	0.98	0.98
SC	**-42.76**	-42.29	-42.18	-28.14							-34.68	-34.16	-33.99	**-35.67**						
SC_3	**0.42**	1.15	1.97	13.38							**0.46**	1.22	2.03	2.77						
Convergence (%)	100%	100%	100%	0%							100%	100%	100%	6%						

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9 and Hor: 12 account for one-, two- three- six-, nine- and twelve-month-ahead forecast, respectively. The RelMSFE is the ratio of the MSFE of a given model against the MSFE of the AR(1) model. VARMA models denoted by $\mathbf{p}_{100}$, $\mathbf{p}_{110}$ and $\mathbf{p}_{111}$ have Kronecker indices equal to $(1,0,0,0,....,0)'$, $(1,1,0,0,....,0)'$, $(1,1,1,0,....,0)'$, respectively. $\mathbf{p}_{HK}$ is a VARMA model with Kronecker indices obtained using the Hannan-Kavalieris procedure on the first rolling window. All VARMA models are estimated using the IOLS algorithm. FM_{IC_3} and FM_{SC} are factor models with number of factors defined using the IC_3 and the SC criteria, where the maximum number of factors is equal to 4. We follow Stock and Watson (2002) and set the maximum lag length for the factor and the variable of interest equal to 3 and 6, respectively. VAR is the VAR(p^*) model with lag length p^* defined by the AIC criterion, with the maximum lag length equal to 8. $BVAR_{SC}$, $BVAR_{0.2}$ and $BVAR_{opt}$ are Bayesian VAR models with the normal inverted Wishart prior which reproduces the principles of the Minnesota-type prior. The three sets of results differ in the way the hyperparameter φ is chosen. $BVAR_{SC}$ adopts φ which minimizes the Schwarz criterion, $BVAR_{0.2}$ uses $\varphi = 0.2$, while $BVAR_{opt}$ is computed with the hyperparameter which minimizes, on every rolling window, the in-sample average of the one-step-ahead root mean squared forecast error of the K variables. SC_K and SC_3 account for the Schwarz criteria computed using the $(K \times K)$ covariance matrix of the residuals and its (3×3) upper block, respectively. The symbols *, **, and *** denote rejection, at the 10%, 5%, and 1% levels, of the null hypothesis of equal predictive accuracy according to the Diebold and Mariano (1995) test.

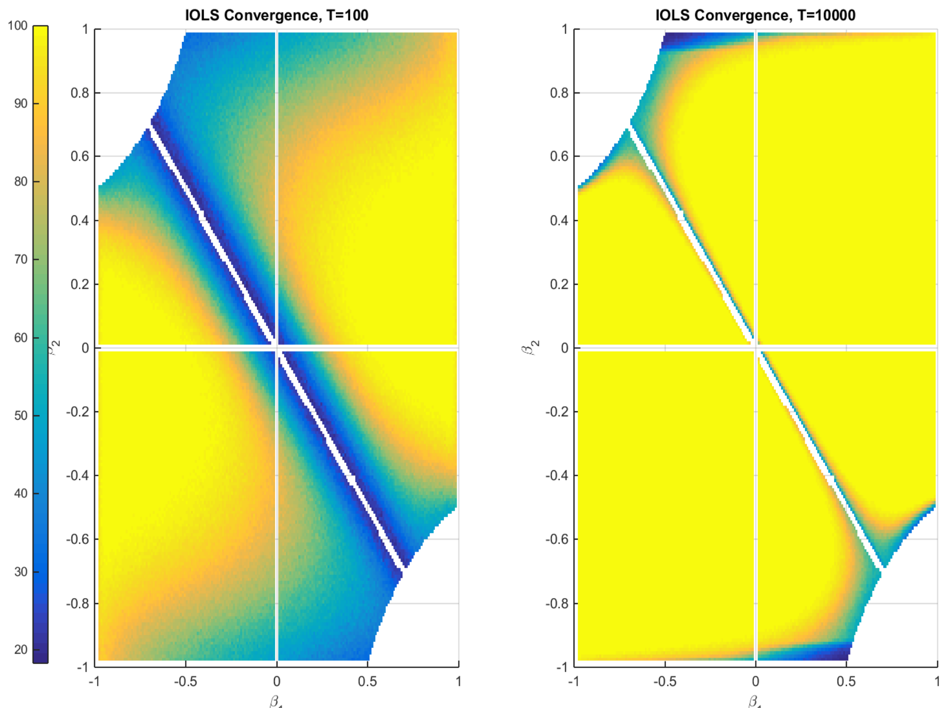
S. 7 Figures

Figure S.2: Maximum Eigenvalue of $V(\beta)$



We plot $|\lambda_2| = \left|\frac{\beta_1\beta_2}{(1+\beta_1\beta_2)}\right|$ computed using different combinations of β_1 and β_2 such that Assumption A.1 is satisfied, i.e., the model is stable, invertible and $\beta_1 \neq \beta_2$. For viewing purposes, we truncate the parameter interval in this analysis such that $\beta_1 = [-0.980, 0.980]$ and $\beta_2 = [-0.980, 0.980]$. The grid is fixed in 0.001.

Figure S.3: Convergence Rates (%)



We plot convergence rates (%) for the IOLS estimator considering the ARMA(1,1) specification considering sample sizes of $T = 100$ and $T = 10,000$. We simulate ARMA(1,1) processes using different combinations of β_1 and β_2 such that Assumption A.1 and Lemma 1 are satisfied, i.e., the model is stable, invertible, $\beta_1 \neq \beta_2$, and $\left|\frac{\beta_1\beta_2}{1+\beta_1\beta_2}\right| < 1$. Convergence rates for each $\beta = (\beta_1, \beta_2)'$ are calculated using 1000 replications. The grid is fixed in 0.01, and the parameters are restricted to the interval $[-0.99, 0.99]$.

References

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics 25*, 71–92.

Brockwell, P. J. and R. A. Davis (1987). *Time Series: Theory and Methods.* Springer-Verlag New York.

Carriero, A., G. Kapetanios, and M. Marcellino (2011). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics 26*, 735–761.

Davidson, J. (1994). *Stochastic Limit Theory. An Introduction for Econometricians.* Oxford University Press.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*, 253–263.

Doan, T., R. Litterman, and C. Sims (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews 3*(1), 1–100.

Dominitz, J. and R. Sherman (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory 21*, 838–863.

Dufour, J.-M. and T. Jouini (2014). Asymptotic distributions for quasi-efficient estimators in echelon varma models. *Computational Statistics & Data Analysis 73*, 69–86.

Dufour, J.-M. and D. Pelletier (2014). Practical methods for modelling weak varma processes: identification, estimation and specification with a macroeconomic application. Manuscript.

Francq, C. and J.-M. Zakoian (1998). Estimating linear representations of nonlinear processes. *Journal of Statistical Planning and Inference 68*, 145–165.

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

Hannan, E. J. and L. Kavalieris (1984a, Aug). A method for autoregressive-moving average estimation. *Biometrika 71*(2), 273–280.

Hannan, E. J. and L. Kavalieris (1984b, Sep). Multivariate linear time series models". advances in applied probability. *Advances in Applied Probability 16*(3), 492–561.

Hannan, E. J. and J. Rissanen (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika 69*, 81–94.

Ibragimov, A. (1962). Some limit theorems for stationary processes. *Theory of Probability and its Applications 7*, 349–382.

Kadiyala, K. R. and S. Karlsson (1997). Numerical methods for estimation and inference in bayesian var-models. *Journal of Applied Econometrics 12*, 99–132.

Kascha, C. (2012). A comparison of estimation methods for vector autoregressive moving-average models. *Econometric Reviews 31*, 297–324.

Koreisha, S. and T. Pukkila (1990). A generalized least squares approach for estimation of autoregressive moving-average models. *Journal of Time Series Analysis 11*, 139–151.

Koreisha, S. and T. Pukkila (2004). The specification of vector autoregressive monving average models. *Journal of Statistical Computation and Simulation 74*, 547–565.

Litterman, R. B. (1986). Forecasting with bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics 4*(1), pp. 25–38.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer- Verlag.

Lütkepohl, H. and D. S. Poskitt (1996). Specification of echelon-form varma models. *Journal of Business & Economic Statistics 14*, 69–79.

Newey, W. K. and D. McFadden (1994, January). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4 of *Handbook of Econometrics*, Chapter 36, pp. 2111–2245. Elsevier.

Romano, J. P. and L. A. Thombs (1996). Inference for autocorrelations under weak assumptions. *Journal of the American Statistical Association 91*, 590–600.

Sims, C. A. and T. Zha (1998). Bayesian methods for dynamic multivariate models. *International Economic Review 39*, 949–968.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Research Papers 2013

2014-20: Martyna Marczak and Tommaso Proietti: Outlier Detection in Structural Time Series Models: the Indicator Saturation Approach

2014-21: Mikkel Bennedsen, Asger Lunde and Mikko S. Pakkanen: Discretization of Lévy semistationary processes with application to estimation

2014-22: Giuseppe Cavaliere, Morten Ørregaard Nielsen and A.M. Robert Taylor: Bootstrap Score Tests for Fractional Integration in Heteroskedastic ARFIMA Models, with an Application to Price Dynamics in Commodity Spot and Futures Markets

2014-23: Maggie E. C. Jones, Morten Ørregaard Nielsen and Michael Ksawery Popiel: A fractionally cointegrated VAR analysis of economic voting and political support

2014-24: Sepideh Dolatabadim, Morten Ørregaard Nielsen and Ke Xu: A fractionally cointegrated VAR analysis of price discovery in commodity futures markets

2014-25: Matias D. Cattaneo and Michael Jansson: Bootstrapping Kernel-Based Semiparametric Estimators

2014-26: Markku Lanne, Jani Luoto and Henri Nyberg: Is the Quantity Theory of Money Useful in Forecasting U.S. Inflation?

2014-27: Massimiliano Caporin, Eduardo Rossi and Paolo Santucci de Magistris: Volatility jumps and their economic determinants

2014-28: Tom Engsted: Fama on bubbles

2014-29: Massimiliano Caporin, Eduardo Rossi and Paolo Santucci de Magistris: Chasing volatility - A persistent multiplicative error model with jumps

2014-30: Michael Creel and Dennis Kristensen: ABC of SV: Limited Information Likelihood Inference in Stochastic Volatility Jump-Diffusion Models

2014-31: Peter Christoffersen, Asger Lunde and Kasper V. Olesen: Factor Structure in Commodity Futures Return and Volatility

2014-32: Ulrich Hounyo: The wild tapered block bootstrap

2014-33: Massimiliano Caporin, Luca Corazzini and Michele Costola: Measuring the Behavioral Component of Financial Fluctuations: An Analysis Based on the S&P 500

2014-34: Morten Ørregaard Nielsen: Asymptotics for the conditional-sum-of-squares estimator in multivariate fractional time series models

2014-35: Ulrich Hounyo: Bootstrapping integrated covariance matrix estimators in noisy jump-diffusion models with non-synchronous trading

2014-36: Mehmet Caner and Anders Bredahl Kock: Asymptotically Honest Confidence Regions for High Dimensional

2014-37: Gustavo Fruet Dias and George Kapetanios: Estimation and Forecasting in Vector Autoregressive Moving Average Models for Rich Datasets